SHORT TERM EX-ANTE LOAD FORECASTING

by

Yike Li

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Infrastructure and Environmental Systems

Charlotte

2022

Approved by:

_____
Dr. Tao Hong

_____
Dr. Pu Wang

_____
Dr. Badrul Chowdhury

_____
Dr. Jy Wu

ABSTRACT

YIKE LI. Short Term Ex-ante Load Forecasting.
(Under the direction of DR. TAO HONG)

Short-term load forecasting (STLF) is a conventional process at power companies to serve for better decision-making in their daily operations. Weather factors play a key role in STLF. In practice, an online STLF system typically requires the use of weather forecasts as input when projecting the future load, with associated weather forecast errors. This type of forecasting is known as ex-ante forecasting. Nevertheless, most existing academic literature developed load forecasting techniques under the ex-post forecasting settings, where the actual weather information is used in the forecast period. Meanwhile, the robustness of STLF models to real weather forecast errors has rarely been studied in the literature. The gap between the practice and the research study is often due to the shortage of historical weather forecasts. In this research, we aim to close this gap by proposing two new frameworks to select better models in short-term ex-ante load forecasting. Compared to the conventional research which focuses on ex-post load forecast accuracy in the model development, both frameworks consider the impact of real temperature forecast errors and are better fitted to field practices.

The effectiveness of the proposed frameworks is confirmed using an empirical case study at a medium-sized US utility with load data from multiple supply areas and real temperature forecasts. Compared to a state-of-the-art benchmark that uses the historical ex-post load forecast accuracy for model selection, the first framework leads to 2.4% improved accuracy on average. A further study among the weather sensitive hours (i.e., the hours when a smaller error in the temperature forecast may lead to a greater inaccuracy in the

load forecast) suggests that the first framework outperforms the benchmark by 3.1% on average. The second framework based on temperature forecast error prediction improves the accuracy of the first framework by 7.4% for the hours with worse predicted temperature forecast accuracy. Overall, the second framework leads to an average of 0.8% improvement over the first framework and 3.9% improvement over the benchmark among the weather sensitive hours.

DEDICATION

To My Family.

ACKNOWLEDGEMENTS

First of all, I want to express my sincere thankfulness to Dr. Tao Hong. I first met Dr. Hong back in the Spring of 2011 at my former graduate school. As an invited lecturer in one of my power system courses, he introduced the subject of electric load forecasting and I immediately get hooked. It was not until 2018 that I became a Ph.D. student in his lab at UNCC and found his teaching style to be very unique. In his class, he assigned open forecasting problems to his students and strongly promoted self-teaching. The grading of each assignment was largely based on the in-class ranking of some forecasting tasks. To receive better grades and maintain self-esteem, I dug into academic papers each week and tirelessly tried out different techniques to improve my forecast. After attending his first class in Fall 2018, I tasted the sweetness of self-teaching and took four other courses under him. At the time I started on my dissertation experiments in 2020, I realize that I have been equipped with most of the necessary knowledge and a box of forecasting techniques to try out and solve the problem. As the Chinese proverb mentions, “Teach a man to fish and you feed him for a lifetime”. Dr. Hong has no doubt been the best mentor I have ever met: he was imperceptibly helping every single graduate student in his lab to become a researcher who can solve real-world problems independently. I would like to express my sincere gratitude to Dr. Hong for his guidance throughout my Ph.D. program.

Secondly, I want to thank the three committee members, Dr. Pu Wang, Dr. Badrul Chowdhury, and Dr. Jy Wu for their generous offering of time and guidance to greatly improve the presentation of this dissertation.

Thirdly, I also want to thank the group members at the Big Data Energy Analytics Laboratory, Vinayak Sharma, Shreyashi Shukla, and Masoud Sobhani. They provided generous help and support for my research.

Moreover, I want to extend my gratitude to the financial support, technical support, and advice from NCEMC. The experiments of this dissertation would not be done without the support from the team members at NCEMC.

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.

Yike Li
Charlotte, NC
Oct 2022

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS

AMI	advanced metering infrastructure
GEFCom2012	Global Energy Forecasting Competition 2012
GEFCom2014	Global Energy Forecasting Competition 2014
GEFCom2017	Global Energy Forecasting Competition 2017
MAPE	mean absolute percentage error
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
PLF	probabilistic load forecasting
MLR	multiple linear regression
DRM	dynamic regression model
ANN	artificial neural network model
RF	Random forecast model
M0	model selection framework based on ex-post forecast accuracy
M1	model selection framework based on ex-ante forecast accuracy
M2	model selection framework based on temperature forecast quality prediction

CHAPTER 1: INTRODUCTION

1.1. Electric Load Forecasting

Electric load forecasting has been a conventional procedure used by power companies to forecast future energy consumption. For decades, power utilities have been upgrading their load forecasting modules to support better decision-making in operations, planning, and maintenances. In recent years, the enormous amount of penetration from renewable energy resources and unprecedented public events such as the spread of the COVID-19 pandemic impose a significant challenge to maintain the equilibrium of supply and demand. Obtaining accurate load forecast nowadays has become increasingly challenging and critical in the era of big data to maintain the stability and reliability of the power grid. In addition, these load forecasts serve as a vital input to other business entities such as regulatory commissions, banks, trading firms, insurance companies, and big commercial companies (Hong & Fan, 2016).

Classification of the load forecasting procedure often depends on the use of the forecast. The industrial and research community often group the load forecasting procedure into categories based on how far the load is projected into the future. While there has not been a universal agreement, (Hong & Fan, 2016) grouped the load forecasting procedure into four categories: very short-term load forecasting (VSTLF), short-term load forecasting (STLF), medium-term load forecasting (MTLF), and long-term load forecasting (LTLF). The cut-off horizons for the four categories are one day, two weeks, and three years, respectively.

The fundamental objective of VSTLF and STLF is to facilitate a real-time balancing between power generation and demand to deliver safe and stable power to the end-users in a financially effective manner. VSTLF and STLF typically provide hourly or more granular level load forecasts and are the cornerstones for smooth day-to-day operations. These load forecasts serve as an essential input for the hour- and day-ahead scheduling, unit commitment, day-ahead energy trading, demand-side management (DSM), and so forth. For instance, a more accurate day-ahead hourly load forecast results in more optimized decisions made in generation unit commitment, economic dispatching, day-ahead energy purchasing, load shedding, and demand side load control. This leads to reduced operational costs while committing to the stability and reliability of the power system.

MTLF and LTLF are vital for longer-term planning and decision-making. These forecasts can typically provide the weekly and monthly peak and valley load information as well as the demand growth in the long run. More accurate medium- and long-term load forecasts are particularly useful for more optimal decisions made in long-term energy purchasing, transmission & distribution network maintenance and expansion, DSM program planning, medium- and long-term revenue projections, and so forth.

The load usage is known to be time-dependent and often displays seasonal patterns. This is impacted by multiple factors. First, the load is largely driven by human activities. Thus, the load pattern varies from each hour of the day, and each day of the week. Besides, the load often displays seasonality in the annual resolution. On the one hand, the load is known to be climate-driven and the climate itself has an annual seasonality due to the Earth's revolution around the Sun. On the other hand, human routines at different months

of the year also play a role in the annual seasonality of load. Hence, calendar variables such as the hours of a day, days of a week, and months of a year are frequently used in load forecasting models.

Weather and climate events can have influential impacts on load usage. Weather is linked to a state of the atmosphere, which often refers to the day-to-day temperature, humidity, wind speed, rainfall, and other atmospheric conditions over a short period of time (e.g., up to one to two weeks). Hence, weather can have a key impact on load usage within a short term. On the other hand, climate refers to a long-term pattern of weather, which could be an average of the aforementioned atmospheric conditions and variations over a longer period of time (e.g., 30 years) within a geographic region. In this case, the long-term variations in climate can affect load usage in the long run.

Weather variables such as ambient temperature, relative humidity, wind speed, solar irradiance, and cloud cover have been widely studied in STLF literature. Compared to the rest, temperature has been the most widely used weather variable due to two reasons. First, compared to other weather variables such as relative humidity and wind speed, empirical case studies have shown that temperature has a stronger correlation to the load (refer to (Xie et al., 2018) and (Xie & Hong, 2017), respectively). Second, the short-term temperature forecasts nowadays are pretty accurate. In contrast, other weather variables such as relative humidity, wind speed, and cloud cover are not as predictable.

The salient correlation between temperature and load is largely due to the heating and cooling needs, which is especially the case for regions with high electrification rates and prevalent installations of electric air conditioners and heaters. During summer, people turn on air conditioning while during winter, people turn on heaters that can be partially or

fully powered by electricity. The rest weather variables such as relative humidity, wind speed, solar irradiance, and cloud cover, can have diverse impacts on load usage. Relative humidity along with the ambient temperature is tied directly to human comfort. Wind speed fastens the water evaporation that cools down the surface of buildings and the human body. Besides, the increased penetration of wind power resources, whose output is largely driven by wind speed, has a direct impact on the regional load. Similarly, solar irradiance and cloud cover have a joint impact on heating the body of premises, as well as the PV power generation. Compared to temperature, these weather variables are not as predictable. Hence, their usage is limited to a relatively short forecast horizon.

Compared to weather factors which trigger more instant disturbances on load usage, economic development and urban expansion may affect the long-term growth trend of the load. Better economic health stimulates economic activities, while urban development and expansion are tied to population growth and the variation in electricity consumption. Hence, macroeconomic indicators such as Gross State Product (GSP), employment rate, and land-use information are often used in medium- and long-term load forecasting.

Depending on what is assumed to be known when generating a forecast, electric load forecasting can be grouped into ex-ante forecasting and ex-post forecasting. Ex-ante is a Latin phrase meaning “before the event”. Ex-ante forecasts are made by solely using the available predictors' information preceding the forecast origin. For instance, to generate the day-ahead ex-ante load forecast using temperature as a predictor, the day-ahead temperature forecast is assumed to be known (and often it is, with forecast errors) and used as input to generate the load forecast. The opposite of ex-ante is ex-post, meaning “after the event”. Ex-post forecasts assume the actual predictors' information is known beyond

the forecast origin. Referring to the day-ahead load forecast example, generating an ex-post forecast assumes perfect knowledge of day-ahead temperature information through the forecast horizon (which is surreal). In this case, the actual temperature readings will be used to generate the ex-post forecasts.

In practice, people use ex-ante forecasts to make decisions. For instance, when generating short-term load forecasts using weather information, power utilities follow a typical procedure as shown in *FIGURE 1*. The load and weather history are first used for the model training process. After a fitted model is in place, the weather forecast through the forecast horizon will be used for model inferencing and producing the final forecast. By evaluating the forecast output under the ex-ante forecasting settings, we get to understand a model's true forecast performance in practice. Nonetheless, it is worth noting that the ex-post forecasts are still useful as they are conventionally evaluated to explore the properties of forecasting models. By comparing the ex-ante and ex-post forecast performance, forecasters get to know whether the forecast inaccuracy is mainly caused by the poor modeling structure or the forecast errors in the predictors.

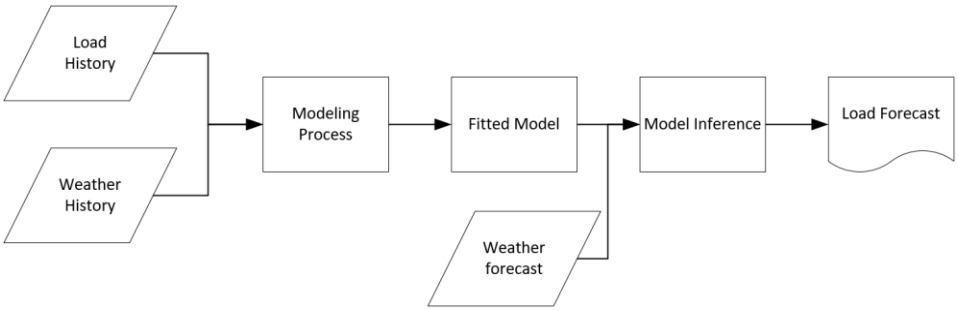

FIGURE 1: A typical STLF procedure using weather information

The quality of weather history is crucial when developing a load forecasting model. The electricity consumption of interest is often an aggregation across multiple geographic

regions, where the weather measurements from multiple weather stations are often available. Due to errors or equipment outages, measurements at certain weather stations may raise data quality concerns for power companies. In addition, measurements at a single weather station can only reveal the weather conditions within a limited geographic region, which may not suffice to describe the load condition of a larger service territory. To address this issue, weather station selection (WSS) methods have been introduced in recent years to select the most relevant weather stations for load forecasting needs.

In real operations, the quality of a weather forecast ties directly to the quality of a load forecast. This is because the load forecasting models typically learn the relationship between the historical weather information and load (as shown in *FIGURE 1*). When predicting future load using weather forecast as input, the associated weather forecast errors, either positive or negative, can add another layer of uncertainty to the load forecast output and eventually degrade the load forecast accuracy. Despite the steady improvement in weather forecast accuracy, weather forecast errors can never be wiped out due to the unforeseeable nature of weather. As the ultimate goal is to have better load forecasting models under the operational (ex-ante) forecasting settings, the impact of weather forecast errors needs to be considered when building and testing the load forecasting models.

Although weather information plays a key role in ex-ante load forecasting, the load forecasts may be generated without using any weather variables. For very short-term load forecasting, such as from minutes to a few hours ahead, the load within an imminent future may be inferred directly from its latest history, where the weather impact on the load is assumed to be consistent. On the other hand, a longer-term load forecasting from hours to days ahead typically requires a weather forecast as input to more precisely capture the

weather impact on the load. As the weather forecast often becomes inaccurate beyond 10 days, simulated weather scenarios based on weather history can be used instead to infer future load scenarios for medium- and long-term load forecasting (e.g., (Hong, Wilson, et al., 2014)).

In this research, we use operational day-ahead temperature forecast data to investigate the ex-ante forecast accuracy, which will be detailed in Chapter 4. In Chapter 5, we propose a new model selection framework aiming to select more robust load forecasting models in the ex-ante forecasting settings and demonstrate the contribution of temperature forecast error to the ex-ante load forecast error. As the load forecasting performance can be impacted by the levels of temperature forecast accuracy (refer to (Methaprayoon et al., 2007), (Segarra et al., 2019), and (Chitalia et al., 2020) reviewed in Section 2.2), we predict the quality of day-ahead temperature forecasts in Chapter 6. Thereafter, we leverage the temperature forecast quality prediction and propose another model selection framework that further improves the load forecast accuracy in Chapter 7.

1.2. Weather Forecasting

Weather forecasting refers to the prediction of the state of the atmosphere for a given location and altitude. Ancient weather forecasting activities can be traced back to over 2000 years ago. Back then, the ways of producing weather forecasts were limited to linking the patterns between sky observations and future weather. It was not until the 1860s that the synoptic weather reports were released in the U.S. and Western Europe, which was largely attributable to the birth of the electric telegraph that enabled rapid dissemination of weather observations to construct a synoptic weather map. With the advent of computer and computer simulations, the first moderately successful numerical

weather prediction (NWP) was released in the 1950s, based on a simple barotropic model (Charney et al., 1950).

Over decades, the accuracy of weather forecasts receives steady improvement. Promoted by the more powerful computers, more complex and realistic atmospheric models were applied to capture the precise state of the atmosphere and oceans. Meanwhile, affordable computing resources enable these models to run at higher resolutions, both horizontally and vertically, for a more accurate representation of the atmospheric physical processes. Along with more observational data from the satellites, weather radars, instruments carried on aircraft and ships, and so forth, as well as more advanced data assimilation techniques, weather forecasting models nowadays have a better grasp of the current state of the atmosphere when depicting the future. Nevertheless, the atmosphere is a highly chaotic system. Recent studies showed that the improvement in weather forecast accuracy slowed down lately as it is approaching the practical predictability of the atmosphere (e.g., (Hoffman et al., 2018) and (Zhang et al., 2019)). Results suggested that the recent NWP forecasts for mid-latitude weather can remain skillful for up to 10 days (Zhang et al., 2019).

There are three main approaches to generating a weather forecast, namely the empirical approach, the dynamical approach, and the statistical approach. The empirical approach infers the future weather condition based on similar weather situations in the past, i.e., the occurrences of analogues. This approach may work well when we have a sufficiently large pool of recorded analogues. The dynamical approach, or so-called numerical weather prediction, is the most widely used approach which generates weather forecasts by solving a set of partial differential equations. First, the equations are

initialized by weather measurements processed via data assimilation techniques, which accommodate the initial conditions to resemble the reality of the atmosphere and minimize the initialization error. Next, the equations are solved iteratively using numerical methods, and the solution is used to determine the next state of the atmosphere. The abovementioned steps can be executed recursively to generate forecasts for a longer lead time (the predicted state of the atmosphere can be used to initialize the model, and the solution of which is then used to determine a further future state of the atmosphere), while the forecast performance may degrade progressively due to the error accumulation at each forecast time step. The NWP systems nowadays usually use a combination of the first two approaches. Due to the data volume involved and the computational intensity of modeling large-scale weather conditions, forecasts from NWP systems cannot be refreshed more frequently. In fact, the time cost of running the models and disseminating the forecast may lead to an already stale forecast when reaching the end-users. In recent decades, statistical approaches have shown effectiveness in performing short-term weather forecasting. Based on recent observations, common approaches such as time series methods and neural networks can generate short-term weather forecasts that are more useful to end-users for near real-time operational needs.

Weather forecast output can be in the form of a deterministic forecast or an ensemble forecast. Traditionally, the deterministic weather forecast is generated using the most likely initial conditions, and the forecast conveys the most likely weather condition in the future. However, due to the uncertainty of the initial conditions or weather forecasting model assumptions, or both, the future weather can have an infinite number of possibilities. Under the ensemble weather forecasting setting, a weather forecast model

will run simulations up to 100 times, each with perturbed initial conditions or model assumptions. If most simulation outputs are very close and form a narrow spread, the weather situation is considered to be highly predictable and the confidence in this prediction is high. Conversely, if these simulation outputs are very different (forming a wider spread), the weather situation is considered to be less predictable, with less confidence that this prediction is accurate. Such an approach to generating the ensemble forecast and inferring the uncertainty of a weather forecast has been accepted and widely used since the 1980s by the meteorological community. Although generating ensemble forecasts can be very computationally costly even with supercomputers, the projection of future weather uncertainties can be particularly useful when the forecast accuracy gets worse at a longer lead time.

Weather forecasting plays a vital role in influencing human daily activities and decision-making. Most industries are weather-sensitive, such as utilities (Teisberg et al., 2005), agriculture (Hansen, 2002), and transportation (Thornes & Stephenson, 2001). Utility companies rely on accurate weather forecasts to project future electricity and gas demand for generation planning and energy trading. Agricultural facilities need weather forecasts for crop management such as irrigation planning, pest control, and consolidating fertilizer requirements. Weather forecasts are also a fundamental input to transportation systems to estimate the adverse weather impact on road conditions, railway operations, airplane routes, and ship operations. For these industries, better weather forecasts are critical to ensure safety, efficient operations, and cost savings. Weather forecast evaluation, or more commonly referred to as weather forecast verification in the meteorological field, is an indispensable part of the weather forecasting systems

development. The evaluation process monitors the weather forecast skill over time, verifies the data assimilation pipeline, and facilitates a comparison of the forecast quality between the weather forecasting systems. By demonstrating the skill improvement and the incremental economic value, the evaluation result helps to build credibility and confidence among its end-users.

1.3. Dissertation Organization

The organization of this dissertation is shown in *FIGURE 2*. Chapter 2 presents a literature review of electric load forecasting, ex-ante load forecasting, and weather forecast evaluation approaches. Chapter 3 presents a background of the statistical models, the model evaluation techniques, and the weather station selection process. Chapter 4 introduces the data used in this research and provides some exploratory data analysis. Chapter 5 proposes a new model selection framework based on ex-ante forecast accuracy and discusses the results. Chapter 6 introduces the steps to model the quality of temperature forecasts. Chapter 7 proposes a comprehensive model selection framework based on results in Chapter 6 and compares the load forecast accuracy to the model selection frameworks discussed in Chapter 5. This research concludes in Chapter 8.

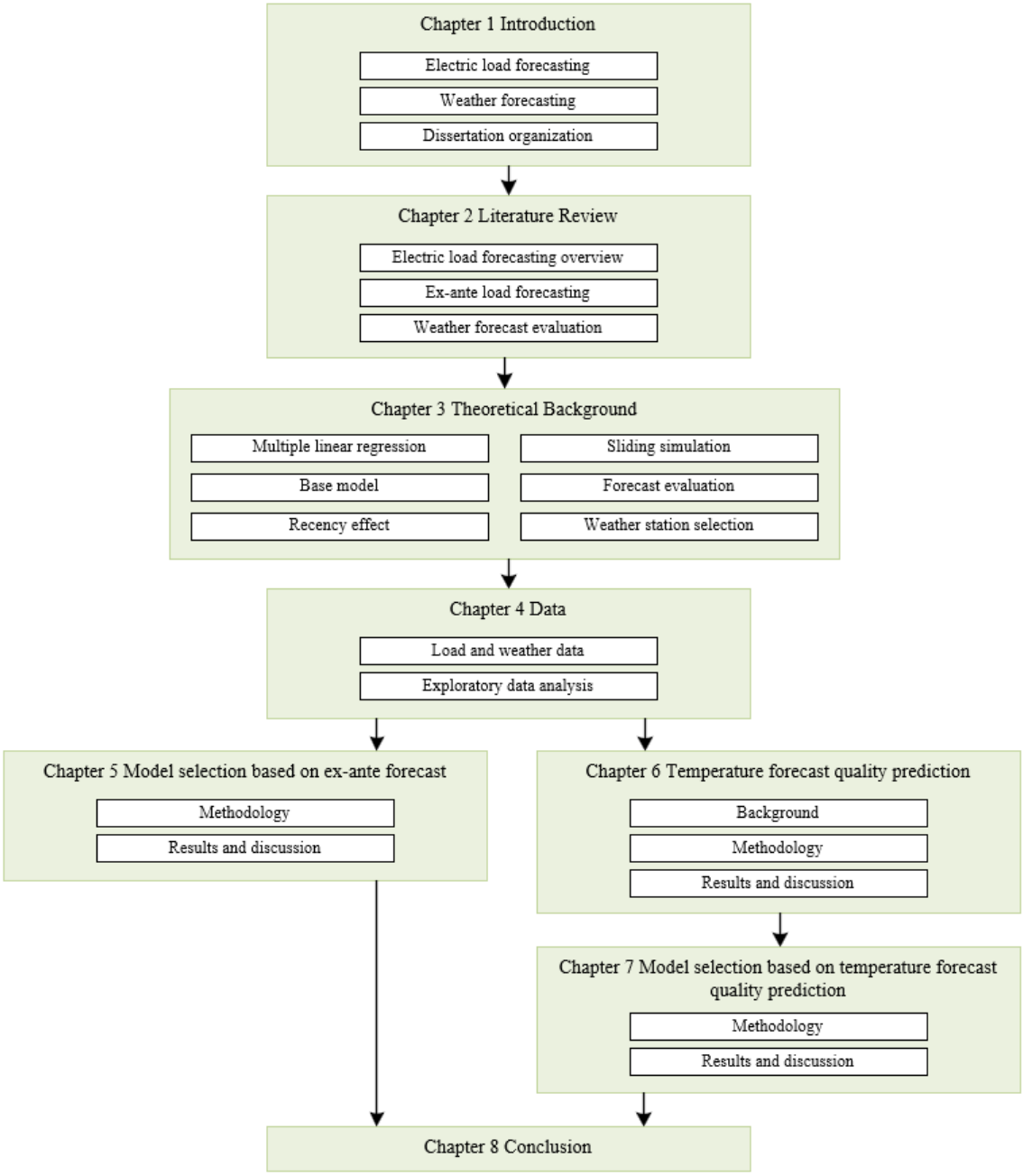

FIGURE 2: Organization of dissertation

CHAPTER 2: LITERATURE REVIEW

2.1. Electric Load Forecasting Overview

Electric load forecasting has received increasing attention from both academia and the utility industry. During the past 40 years, thousands of papers have been published in this field. *FIGURE 3* shows the number of journal articles published in the electric load forecasting field since 1985. The numbers are scraped from Web of Science based on the query: TS=(load forecasting OR electricity consumption forecasting OR electric load forecasting OR energy forecasting). Likely due to the new technologies being adopted and the availability of big data, a sharp increase in the number of publications can be observed during the most recent decade. In this section, we review some notable load forecasting literature based on the following three aspects: 1) the length of the forecast horizon, 2) the load forecasting techniques being explored, and 3) the common variables being used in the load forecasting applications. In the end, we touch on several trending topics in this field.

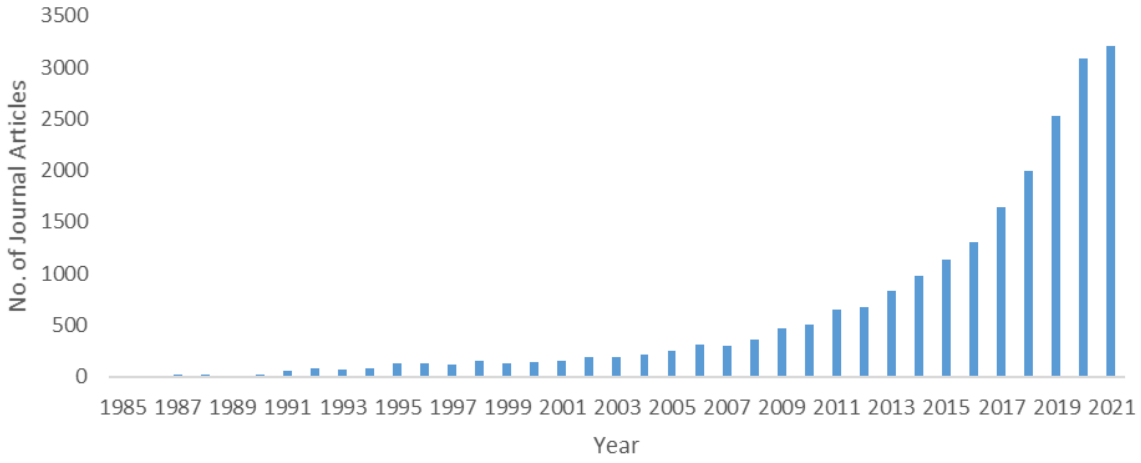

FIGURE 3: Number of journal articles in electric load forecasting based on Web of Science query (1985 – 2021)

As mentioned in Section 1.1, load forecasting can be grouped into four categories: VSTLF, STLF, MTLF, and LTLF. VSTLF can be viewed as a sub-problem of STLF as both processes can take weather forecasts as input in the forecasting period (Luo, Hong, & Yue, 2018). Nevertheless, the weather information can be optional in VSTLF since the load in a near future can be directly inferred from the load in the past (Taylor & McSharry, 2007)(Hong, 2010).

For VSTLF, (Taylor, 2008) evaluated 10 to 30 minutes ahead forecasts of British electricity demand among a few univariate methods plus a multivariate method that used weather forecast as input. Empirical results suggested that the forecast performance deteriorated with a longer lead time for all methods. Compared to the univariate methods, the weather-based multivariate method was not competitive for very short-term prediction, whereas it led to superior performance for lead times longer than four hours. (Mandal et al., 2006) generated one to six hours ahead load forecasts based on load data from Okinawa Electric Power Company in Japan. The proposed algorithm first located similar days by comparing the load and temperature forecast to the past. Then, the average load of similar days from the past served as input to an ANN model to produce the final forecast. (Methaprayoon et al., 2007) applied an ANN model to generate hour-ahead and day-ahead forecasts for Western Farmers Electric Cooperative. To improve the quality of the input temperature forecast, the authors trained another ANN model which used source temperature forecasts as input and tried to minimize the difference between the model output and actual temperature. The results showed that this extra step improved the load forecast accuracy upon the usage of any individual temperature forecast. When forecasting the heating and cooling load of an office building, (Zhao & Liu, 2018) observed a better

weather forecast accuracy within 1-hour, 2-hour, and 3-hour ahead compared to the 24-hour ahead weather forecast. Such a fact contributed to a lower load forecasting error within a shorter lead time. As the preceding hours' load variable was commonly used in online VSTLF systems, (Luo, Hong, & Yue, 2018) proposed a model-based anomaly detection method to detect and cleanse corrupted load data.

There is rich literature in STLF. (Hong, 2010) presented a comprehensive review of this field for the recent half century until 2010. In this dissertation, the author set the foundation for a modern load forecasting module using multiple linear regression (MLR) models for week-ahead hourly load forecasting. A base model was constructed by exploring the relationship between load, temperature, and calendar variables. The proposed base model can easily be extended for VSTLF by including the preceding hour load, and for MTLF/LTLF by including econometric variables. Besides, a few customizations to the model, such as including the recency effect, weekend effect, and holiday effect enhanced the predicting power significantly on top of the original model. The base model is known as Tao's Vanilla Benchmark model and has been extensively studied in recent STLF literature, such as (Hong, Wilson, et al., 2014), (P. Wang et al., 2016), (Xie & Hong, 2017), and (Xie et al., 2018). Given its accuracy and computational efficiency, the base model has been used as a benchmark model in recent load forecasting competitions including GEFCom2012 (Hong, Pinson, et al., 2014), 2014 (Hong et al., 2016), and 2017 (Hong et al., 2019).

Regarding the number of research papers getting published, MTLF and LTLF are much less popular topics compared to STLF. Nevertheless, MTLF and LTLF are equally important as STLF. (Khuntia et al., 2016) presents a comprehensive review on defining

and classifying various load forecasting techniques for MTLF and LTLF. Compared to STLF, which emphasizes model fitting on existing datasets without the need of understanding the way a power system works, the authors pointed out that MTLF/LTLF relied more on a forecaster's experience with the power system itself, as well as the understanding of how the economy and technological changes may affect the electricity market and the load usage. As the temperature cannot be predicted accurately in the long run, (Hong, Wilson, et al., 2014) incorporated economic scenarios and created weather scenarios from past weather history for the LTLF of a US utility. The model outputted a probabilistic load forecast which represented diverse scenarios of the future load.

An extensive variety of techniques have been explored in the load forecasting literature. The mainstream techniques can roughly be grouped into two categories: statistical methods such as time series analysis and regression analysis, and artificial intelligence (AI) methods such as artificial neural networks (ANN) and gradient boosting machines (GBM). *TABLE 1* presents some notable methods and the related literature.

TABLE 1: Notable methods and literature in the electric load forecasting field

Method	Papers
ARIMA	(Amjady, 2001)(Weron, 2006)
Exponential smoothing	(Hyndman et al., 2008)(Taylor, 2008)
MLR	(Papalexopoulos & Hesterberg, 1990)(Hong, 2010)(P. Wang et al., 2016)
Semi-parametric	(Fan & Hyndman, 2012)(Goude et al., 2014)
ANN	(Hippert et al., 2001)
Support vector machine	(B. J. Chen et al., 2004)
GBM	(Taieb & Hyndman, 2014)
Deep learning	(K. Chen et al., 2019)

Statistical methods such as univariate models have been popular approaches in the 2000s literature. The family of time series methods such as ARMA models (Weron, 2006)

and exponential smoothing methods (Taylor & McSharry, 2007)(Taylor, 2008) was widely studied and compared. These univariate methods commonly learn from the past load history without the knowledge of exogenous variables such as weather. Therefore, they have lower data requirements than the other widely used techniques such as MLR and ANN (Hong & Fan, 2016). Nevertheless, empirical results in (Taylor & McSharry, 2007) and (Taylor, 2008) showed that these univariate methods only led to promising forecasting performance for VSTLF with the lead times up to about four to six hours. At a longer lead time, their forecasting performances dropped significantly with the forecast error accumulation at each forecast time step and were eventually surpassed by other multivariate methods.

On the other hand, MLR models are frequently used in the literature for both STLF and LTLF. These models require statistical knowledge to establish a functional form between the input variables (e.g., weather and calendar variables), and the output variable (e.g., load or some transformation of the load). Compared to the "black-box" models such as ANN, the parameters of MLR models are much easier to interpret. (Papalexopoulos & Hesterberg, 1990) set a solid ground in employing MLR analysis for STLF. The proposed approach modeled temperature by using heating and cooling degree functions, holiday effect by using binary variables, and estimated parameters using weighted least squares. Using Pacific Gas and Electric Company's (PG&E) data, the proposed method outperformed the existing approach at PG&E for 24-h ahead peak and hourly load forecasts. (Hong, 2010) proposed a series of MLR models for week-ahead hourly load forecasting. The proposed approaches emphasized the cross effects between weather and calendar variables and addressed special effects in the load forecasting models such as the weekend

effect, holiday effect, and recency effect. (Hong, Wilson, et al., 2014) introduced a macroeconomic variable to the STLF model in (Hong, 2010) and incorporated economic scenarios and weather scenarios for long term probabilistic load forecast. (P. Wang et al., 2016) extended the base model in (Hong, 2010) and formally investigated the best number of recency effect terms (lagged hourly temperatures and lagged moving average temperatures) using high-performance computing.

AI methods such as ANN and GBM became increasingly popular in recent decades. These techniques typically do not specify an exact functional form between the input and output variables. Instead, the relationship between the two is automatically learned by minimizing a cost function. Despite the ease of effort in the feature engineering stage, the cost of tuning the model's architecture and hyperparameters is nontrivial. Besides, the learned parameters can be difficult to interpret and used to explain the relationship between the input and output variables (so-called the "black-box" models). (Hippert et al., 2001) presented a comprehensive review of the usage of ANN in STLF. (Taieb & Hyndman, 2014) employed a gradient boosting framework to solve the hierarchical load forecasting problem in GEFCom2012, where the univariate penalized regression splines were used as the base learner.

Apart from the abovementioned techniques, other methods were also considered when building load forecasting models, which included but were not limited to fuzzy regression models (Hong & Wang, 2014), support vector machines (B. J. Chen et al., 2004), semi-parametric models (Fan & Hyndman, 2012)(Goude et al., 2014), and the most trending deep learning models (K. Chen et al., 2019).

Various variables have been explored in the load forecasting literature. These variables are used to extract useful information in explaining load variations. The most common variables are weather and calendar variables.

Temperature is the most widely used weather variable due to two aspects. First, it has a strong correlation with weather-sensitive load, as shown in *FIGURE 7*. Second, temperature forecasts are usually pretty accurate within four or five days ahead. As weather forecast is required when inferencing the future load, temperature, which is more predictable than the other weather variables, becomes a more reliable input to load forecasting models. The non-linear relationship between the measured temperature and electricity usage has been modeled in various forms. (Shu Fan et al., 2009) used a piecewise linear function to estimate the correlations between load and temperature for cold and hot days. The separation point was located by maximizing the absolute values of the two correlation coefficients on both segments. Similarly, (Ziel & Liu, 2016) manually chose two separating points, 50 °F and 60 °F, when fitting a piecewise linear function. (Hong, 2010) explored a 3rd order polynomial function to model the nonlinear relationship between temperature and load. The author further introduced cross effects between temperature and calendar variables to capture the varying correlations between temperature and load at diverse temporal levels (different hours of a day and months of a year). Since the temperatures of the preceding hours can impact the current hour's load, (P. Wang et al., 2016) explored the best number of lagged hourly temperatures and lagged moving average temperatures to be included in a regression-based model. Besides, (Chapagain & Kittipiyakul, 2018) converted hourly temperature data to heating degree days (HDD) and

cooling degree days (CDD), and included their 3rd order polynomials as well as their cross effect with the calendar month variable in their STLF model.

Other weather variables have also been explored, such as relative humidity, wind speed, solar irradiation, and rainfall. (Xie et al., 2018) experimented with relative humidity variables in a form of the heat index, a 2nd order polynomial of relative humidity, and some cross effects with temperature, season, and hour of the day. A case study based on a North Carolina utility showed that relative humidity had a stronger correlation with the load during the summer months (June to September). (Xie & Hong, 2017) explored wind speed variables in a form of the wind chill Index, and some cross effects among the wind speed, temperature, and season. A case study based on ISO New England data suggested that wind speed had a stronger correlation with the load from June to August. (Zhou et al., 2008) generated solar irradiance forecast based on existing weather forecast information and the solar irradiance output was used for building load forecasting. (Senjyu et al., 2005) introduced rainfall as a binary variable (-1: rain; 1: no rain) to forecast the next day's load of a utility company in Japan.

As weather inputs are crucial to load forecasting models, the quality of weather variables can largely affect the accuracy and robustness of a load forecasting model. Besides, the data measured at a single weather station may solely reflect the weather conditions within a limited geographic region (Sobhani et al., 2019). When data from multiple weather stations are available, one way to enhance the quality of weather input is through a weather station selection process. (Hong et al., 2015) was the first paper published on this subject. In this paper, the authors proposed a heuristic method based on a greedy algorithm to select a subset of weather stations that was suitable for load

forecasting models. The method produces relatively accurate results and is fast to compute. Due to its simplicity and reproducibility, this method has been cited and compared in many recent literature (e.g., (Xie & Hong, 2016), (Sobhani et al., 2019), and (Neto & Hippert, 2020)) and is currently used by many power companies (Y. Wang et al., 2019). (Moreno-Carbonell et al., 2019) is another notable paper in this field, in which the authors employed a GA (genetic algorithm) based method that resulted in superior accuracy with more computational costs.

The load consumption is known to be time-dependent, and calendar variables can be used to capture the load levels at different time slots. The Gregorian calendar is frequently referred to when building load forecasting models. (Hong, 2010) introduced the hour of the day, day of the week, and month of the year as categorical variables to model the load levels at different temporal categories. The author also explored groupings of similar weekdays and modeled the load around holidays by assigning different days of the week categories to these days. Other papers investigated groupings of consecutive months to model the seasonal variation of the load, such as summer and winter periods (Charlton & Singleton, 2014), four seasons (Hagan & Behr, 1987), or two seasons with a transition period (Lusis et al., 2017). (Sangamwar, 2019) employed heuristic algorithms to explore an optimal grouping of the calendar variables, including the hour of the day, day of the week, and month of the year. Apart from the Gregorian calendar, some works reported using the solar terms (Xie & Hong, 2018) or seasons based on the lunar calendar (K. Chen et al., 2019) in the load forecasting models. Besides the usage of categorical variables, other literature implemented smooth cyclic functions such as cyclic cubic regression

splines (Nedellec et al., 2014)(Gaillard et al., 2016) and Fourier approximation (Taieb & Hyndman, 2014)(Haben et al., 2019) to estimate the calendar-related cyclic pattern of load.

The traditional load forecasting procedure outputs a point forecast, which measures the expected load value at each future time step. In recent years, probabilistic load forecasting (PLF) has received increasing attention. This is due to the fact that PLF can capture future load uncertainty, which is critical for the decision-making of grid operations and maintenances with the integration of renewable energy resources. (Hong & Fan, 2016) provided the first comprehensive review on PLF across all forecast horizons, including the techniques being used, the common PLF applications, and forecast evaluation methods. The two recent energy forecasting competitions, GEFCom2014 and 2017, invited contestants around the world to address the PLF problems. (Hong et al., 2016) and (Hong et al., 2019) summarized the PLF techniques being used in each competition, such as quantile regression (Haben & Giasemidis, 2016)(Liu et al., 2017), temperature scenario simulation (Gaillard et al., 2016), and residual simulation (Xie et al., 2017).

Due to massive installations of metering devices in the power distribution network since the early 2000s, highly granular data become available at diverse power system levels, providing insights into energy usage from a single residential property to a major utility service territory. The massive data spawns a new subject of hierarchical load forecasting (HLF) which aims to provide comprehensive load forecasts at diverse power system hierarchies to support operations and marketing strategies. The HLF literature presents a growing trend in the recent decade. (Shu Fan et al., 2009) developed a day-ahead multi-region load forecasting system for a Midwest power utility in the US. The proposed method explored the optimal region partition under diverse weather and load conditions to achieve

more accurate aggregated load forecasts. To improve load forecast accuracy based on the ISO New England data, (Lai & Hong, 2013) investigated regional grouping using geographical hierarchies and an average of temperature stations. The hierarchical load forecasting track in GEFCom2012 presented contestants with a 2-level HLF problem: to backcast and forecast hourly load for a US utility at both the zonal level (20 zonal series) and the system level (sum of 20 zonal series). (Hong, Pinson, et al., 2014) summarized the challenges of the problem and the techniques being used by the top winning entries, such as multiple linear regression (Charlton & Singleton, 2014), semi-parametric models (Nedellec et al., 2014), and gradient boosting machines (Lloyd, 2014)(Taieb & Hyndman, 2014). The GEFCom2017 escalated the HLF problem to a 3-level (qualifying match, based on zonal and total loads of ISO New England) and 4-level (final match, based on the load of hundreds of delivery point meters for a US utility) PLF problem. (Hong et al., 2019) summarized the techniques being used and found a modest usage of the hierarchy information by the top winning entries.

Load forecast accuracy relies heavily on the quality of load history. As cybersecurity brings a growing concern to diverse business enterprises, data integrity attacks have become an emerging issue in modern load forecasting systems. (Luo, Hong, & Fang, 2018a) investigated four representative load forecasting models under various simulated data integrity attacks. Using the GEFCom2012 data, the authors found that the support vector regression (SVR) model was the most robust model compared to the MLR, ANN, and fuzzy interaction regression (FIR) models. In (Luo, Hong, & Fang, 2018b), the authors demonstrated that the L1 regression model showed much stronger robustness than the other two iteratively re-weighted least squares models. (Luo, Hong, & Yue, 2018)

proposed a real-time anomaly detection method to prevent malicious data attacks on the data acquisition system.

Since early 2020, the spread of the COVID-19 pandemic brought an unprecedented impact on human society. Due to lockdown measures, the electricity demand in Spain decreased by over 13% in a month (Santiago et al., 2021). After state-wise curfew orders, New York and California independent system operators reported a 10% and 12% reduction, respectively, in electricity demand (Alasali et al., 2021). Besides, the morning and evening peaks of homes, hospitals, and the total electricity demand were modified due to changes in customer behaviors. All these changes brought extra uncertainty to the reliable operation of the grid. At this moment, notable papers considering the COVID impact on electric load forecasting are still scarce. (Alasali et al., 2021) analyzed the COVID impact on energy demand trends of three supply areas in Jordan by applying time series decomposition. The authors applied a rolling origin ARIMAX model with a Monte Carlo sampling method to generate PLF. (Tudose et al., 2021) proposed a convolutional neural network based model for the day-ahead forecast of aggregated load in Romanian. The authors modeled the pandemic effects to load consumption by using categorical variables that indicated whether any restrictions were applied during the forecast day. The restrictions were known based on publicly announced governmental decisions, namely a national lockdown, a period of gradual relaxation, or a partial lockdown.

2.2. Ex-ante Load Forecasting

In the load forecasting field, it has been a conventional practice to build and test models in the ex-post forecasting settings, in which the actual weather readings in the forecast period are used. In short-term load forecasting, the temperature can be fairly

predictable, and we may assume that the models selected by the ex-post forecasting can represent the performances in the operational forecasting. For long-term load forecasting, ex-post forecasting can also be used to explore the modeling strategies (Hong, Wilson, et al., 2014). This leads to the fact that the ex-post forecasting performance is typically reported, which enables comparisons with other research (Hong, Wilson, et al., 2014).

The underlying assumption of studying load forecasting models in the ex-post forecasting settings is that, a more accurate model measured based on the ex-post forecasting performance would lead to more accurate ex-ante forecasts in the operational context. Nevertheless, recent literature have shown that this assumption may hold in some case studies (e.g., (Dahl et al., 2018)), but not necessarily in the other ones (e.g., (Sandels et al., 2015) and (Z. Wang et al., 2020)). This is because weather forecast errors can lead to diverse levels of performance degradation in different load forecasting models. Such effects are completely overlooked in the ex-post forecasting settings.

As the ultimate goal is to improve the ex-ante forecast accuracy, historical weather forecasts can be used for forecast evaluation in the model building stage. However, the ex-ante forecasting performance was rarely reported in the literature, which is often due to the lack of historical weather forecast data for the training period (Fay & Ringwood, 2010). It was not until recently that the historical numerical weather prediction (NWP) was made publicly available by the European Centre for Medium-Range Weather Forecasts (ECMWF) (Yang et al., 2022). This dataset contains four years of historical weather forecast history, 14 weather forecast variables, and covers most of the Europe and North America regions. We anticipate seeing more literature being published on investigating the

ex-ante forecast accuracy of load forecasting models using the real weather forecast from this dataset.

As weather forecast often becomes inaccurate beyond 10 days, existing literature using weather forecast variables typically focuses on forecasting the load within a short horizon. In this review, we focus on discussing the short-term ex-ante load forecasting literature as it is directly tied to our research topic. We review these literature based on the following three aspects: 1) the ways ex-ante load forecasts were generated; 2) the existing analyses on the contribution of weather forecast errors to the ex-ante load forecast errors, and 3) the efforts in identifying more robust load forecasting models considering the weather forecast errors.

In the literature, ex-ante load forecasts were most commonly generated using a real weather forecast. Such a forecast could be a point forecast provided by a single (De Felice et al., 2013)(Zhao et al., 2018)(C. Fan et al., 2020) or multiple weather forecast providers (Methaprayoon et al., 2007)(S. Fan et al., 2008), or an ensemble weather forecast from a single provider which consists of multiple future weather scenarios (Taylor & Buizza, 2002).

In (De Felice et al., 2013), the authors used the operational weather forecast from ECMWF (European Centre for Medium-range Weather Forecasts) to produce the ex-ante forecast of regional load in Italy. A performance gap was observed between the ex-ante forecast and the ex-post forecast, which was due to the weather forecast errors within the operational forecast. (Zhao et al., 2018) retrieved weather forecast data from a weather website and forecasted the cooling load of an office building. Instead of using the weather forecast directly, the authors first conducted Monte Carlo (MC) simulations to sample the

errors from the historical weather forecast error distribution. Next, they applied the sampled errors to revise the future weather forecasts. The case study showed that the revised weather forecasts improved load forecast accuracy. Compared to (Zhao et al., 2018), (C. Fan et al., 2020) argued that the probability distribution of the weather forecast errors should be determined based on the changing characteristics of the errors. The authors proposed a similar MC framework to sample weather forecast errors from a two-period probability distribution: one formed by the weather forecast errors from 6:00 AM to 4:00 PM when the outdoor temperature was typically rising, and the other one formed by the errors from 4:00 PM to 6:00 AM when the outdoor temperature was typically dropping. The sampled errors were likewise used to calibrate future weather forecasts.

The temperature forecasts in (Methaprayoon et al., 2007) and (S. Fan et al., 2008) were both retrieved from multiple weather sources. To make better use of the data, different approaches were applied to pre-process these source temperature forecasts. In (Methaprayoon et al., 2007), the authors trained an ANN model which took different temperature forecasts as input and tried to minimize the difference between the output and actual temperature. The proposed module was intended to output a more effective temperature forecast to be used as input for ex-ante load forecasting. The results showed that this extra step led to a better load forecast accuracy than the usage of any individual temperature forecast. Based on another case study in (S. Fan et al., 2008), the authors found that the temperature forecasts generated by the ANN module in (Methaprayoon et al., 2007) sometimes were not satisfactory in practical application. When the performance of the source temperature forecasts varied over time, which is common in the real-time application, the temperature forecast generated by the ANN module can lead to worse load

forecast accuracy than using the source temperature forecast directly. Instead, the authors proposed a more stable solution by combining these source temperature forecasts using adaptive weights. The weight of each source temperature forecast was determined by two factors: one is its average performance in the history compared to the real temperature readings; the other one is its latest performance through an exponential term. The case study showed that the combined temperature forecasts led to better load forecast accuracy as well as lower standard deviation of the load forecast errors. When multiple weather stations are available within or near a service territory, (Hong et al., 2015) demonstrated a heuristic framework to select the most suitable weather stations for load forecasting models. Although the steps of proposed framework used ex-post forecasting performance as the selection criteria, historical weather forecasts can instead be used to select the stations based on the ex-ante forecasting performance.

(Taylor & Buizza, 2002) forecasted the mid-day daily load of England and Wales from one to ten days ahead. They used a weather ensemble prediction generated by ECMWF, which consisted of 51 future weather scenarios for each weather variable, i.e., temperature, wind speed, and cloud cover. The authors applied each weather scenario as input to produce ex-ante load forecasts. Results showed that the average of 51 load forecasts outperformed the forecast produced by traditional point weather forecasts for all the lead times. Besides, the distribution of the load forecasts can provide uncertainty in the future load.

When real weather forecast was unavailable or incomplete for a load forecasting model, the research community mostly generated weather forecasts following one of two ways: 1) build weather forecasting models from past weather history (Charlton & Singleton,

2014)(Lloyd, 2014)(Nedellec et al., 2014)(Kulkarni et al., 2013) or other weather forecast variables (Zhou et al., 2008), or 2) create synthetic weather forecast based on actual weather readings (S. T. Chen et al., 1992)(Park et al., 1993)(Fay & Ringwood, 2010).

In GEFCom2012, weather information during the forecast period was unprovided. Therefore, the contestants need to develop their own weather forecasts if they decided to include weather information in their models. (Charlton & Singleton, 2014) generated temperature forecasts by simply averaging the past time-of-the-year temperatures within a fixed window of 25 days. (Lloyd, 2014) predicted temperature by modeling a smooth trend and daily periodicity of past temperature. The smooth trend was estimated using a local linear regression with a bandwidth of one day, while the daily periodicity portion was modeled using a Gaussian process with a periodic kernel. (Nedellec et al., 2014) explored five different approaches to forecast temperature based on historical temperature, including SARIMA models, semi-parametric models, and a kernel wavelet functional forecast. The authors proceeded with semi-parametric models with SARIMA errors as the temperature forecast led to the lowest load forecast error. Besides the winning entries in GEFCom2012, (Kulkarni et al., 2013) built a spiking neural network based on past weather history to forecast the day-ahead hourly temperature. The weather forecast output was then used to generate the ex-ante load forecast. (Zhou et al., 2008) proposed a grey-box model to forecast the building cooling load, where the raw weather forecast information was retrieved from a local observatory website. However, some critical weather variables such as hourly temperature, relative humidity, and global solar radiation were unavailable from the local source. Therefore, the authors utilized grey models and linear estimators to forecast these weather variables based on existing weather forecast information.

Rather than using real weather forecasts for ex-ante load forecasting, (S. T. Chen et al., 1992) and (Park et al., 1993) created synthetic temperature forecasts by adding Gaussian noises to the actual temperatures. A more recent empirical case study presented in (Zhao et al., 2018) also showed that the real weather forecast errors followed a distribution that was close to normal. On the other hand, (Fay & Ringwood, 2010) examined the error distribution of weather forecasts from the meteorological office of Ireland. The authors found that the weather forecast error displays a serial correlation, meaning that a more complicated structure was necessary to model the pattern of such errors. To create the synthetic temperature forecast for the entire training period, they modeled the temperature forecast error by introducing turning points, a level error, a shape error, and a random error. The model output showed that this method well captured the auto-correlation within the historical temperature forecast errors. Nevertheless, the usage of real weather forecasts is still highly recommended to validate a load forecasting technique. This is because using the synthetic weather forecast may result in a simplified and more predictable problem (Agüera-Pérez et al., 2018), which is non-identical to the problems in practical operations.

The ex-ante load forecast output typically consists of two uncertainties: one is stemmed from the forecasting model itself, while the other one is from the weather forecast erors. Several notable literature investigated the contribution of weather forecast error to the ex-ante load forecast error. (Douglas et al., 1998) compared the ex-ante and ex-post forecasting performance and claimed that a sizable portion of the load forecast error is due to a lack of accuracy in the temperature forecast. (Segarra et al., 2019) noticed a strong positive correlation between the temperature forecast error and the load forecast error. Both

errors generally increased together when reaching a longer forecast horizon. (Methaprayoon et al., 2007) observed significant load forecast errors when temperature forecast errors were high, implying that the performance of their load forecasting model was sensitive to the accuracy of the temperature forecast. The authors subsequently conducted a sensitivity analysis of temperature forecast error to ex-ante load forecast error. The load forecasting performance was simulated by applying a level shift error to the temperature variable from -5 to 5 °F. The simulation results showed that the load forecast accuracy dropped with a larger deviation of the (simulated) temperature forecast from the actual. (Chitalia et al., 2020) conducted a similar sensitivity analysis using Gaussian distribution with the mean of the actual temperature and three standard deviations of 5%, 10%, and 15%. The study showed that different candidate models received different levels of degradation in performance under various levels of temperature forecast errors. The abovementioned case studies provide evidence that the ex-ante load forecast accuracy can be impacted by the levels of temperature forecast accuracy.

Two other literature touched on identifying more robust load forecasting models to weather forecast errors. Possibly due to the lack of weather forecast data, both papers added Gaussian noises to actual weather to simulate ex-ante forecast. (Z. Wang et al., 2020) compared the building load forecasting performance between XGBoost and LSTM. Results showed that LSTM, which was slightly worse than XGBoost under ex-post forecasting, was more robust to synthetic weather forecast errors (temperature and relative humidity). To increase the robustness of a model to the weather forecast errors, the authors tried on using synthetic weather forecasts instead of real weather readings for model training. Both models showed improvement when trained with synthetic weather forecasts.

(Chitalia et al., 2020) compared nine different hybrids of RNNs, where the synthetic temperature forecast was created by adding Gaussian noise with the mean of the actual temperature, and three standard deviations of 5%, 10%, and 15%. Sensitivity analysis was carried out among these models and the LSTM model with the attention mechanism performed the best under all three standard deviations of the Gaussian noise. Besides, the performance ranking across different models changed drastically under different levels of temperature forecast error. Although the above two studies discussed the robustness of load forecasting models using simulated ex-ante forecasts, our case study (introduced in Chapter 4, *FIGURE* 9) showed that the real temperature forecast was not alike random noises injected into the actual temperature. Hence, further investigations are necessary to determine the better load forecasting models in field operations.

The reviewed literature revealed two following directions on improving ex-ante load forecast accuracy: 1) improving the accuracy of weather forecast input to load forecasting models, and 2) improving the robustness of load forecasting models to the weather forecast errors. Some abovementioned literature shed some light on the first direction, while the literature suffers more from a lack of comprehensive study on the second direction, which is to choose more robust load forecasting models under real weather forecasts. In this research, we offer an extensive investigation of the better load forecasting models using real temperature forecasts purchased by a US power utility.

2.3. Weather Forecast Evaluation

There are numerous ways to evaluate the goodness of a weather forecast. From the forecaster's perspective, the goodness of a weather forecast mainly refers to the extent of correspondence between the forecast and observations. On the other hand, the end-user

is more interested in the incremental beneficial outcomes from their decision-making supported by the weather forecast. In the meteorological field, (Murphy, 1993) suggested three types of goodness that jointly define the goodness of a weather forecast: consistency, quality, and value. Consistency relates to the correspondence between the forecast and the forecaster's best judgment, which encourages proper scoring rules being used during forecasting. Quality refers to the correspondence between the forecast and observations. The closer a forecast is to the observations, the better quality it has. Value refers to the incremental economic benefits of the forecasts to the end-user. For instance, (Teisberg et al., 2005) concluded that a 1°F improvement in temperature forecast accuracy was worth about $37 million per year to cut down the operational cost of electricity generation in the U.S. North, South, or West.

In this section, we first review the common approaches to evaluating weather forecast quality. We limit our discussion to the quality of deterministic weather forecasts on continuous variables (e.g., temperature, precipitation amount, solar irradiance, etc.). Evaluation methods related to other forms of the weather forecast, such as categorical variables (e.g., binary events such as tornado vs. no tornado), word descriptive forecast, and ensemble (probabilistic) forecast are not within our scope. Going beyond the weather forecast evaluation, we review a handful of literature that attempted to predict the forecast quality of NWP systems in the end.

Data used for weather forecast evaluation can be divided into reanalysis and operational forecast. Reanalysis data, also known as the "hindcast", is produced by a modern dynamic weather model with fixed initial conditions, physical schemes, and parameterization. It provides a complete picture of possible past weather and climate for

multiple decades at once. In contrast, operational forecast nowadays is generally produced by multiple dynamic weather models with a myriad of ensemble members within each model. The forecast is generated through the forecast horizon (usually within 15 days) for a designated region and disseminated to the end-user iteratively when a newer forecast becomes available. As the sample size (both temporal and spatial) of historical operational forecasts is often limited, the reanalysis data with a larger size is often used to derive the long-term skill trend of an NWP system. Since reanalysis is not equivalent to operational forecasts, several caveats, detailed in (Jolliffe & Stephenson, 2012), remain when using reanalysis data to estimate the skill of an NWP system in the operational context.

In the meteorological community, evaluation methods of the forecast quality can be divided into the measure-oriented approach and the distribution-oriented approach (Murphy, 1993). The measure-oriented approach, such as the mean absolute error (MAE) or the root mean square error (RMSE), measures the overall aspects of the forecast quality. These "one-number summary" error metrics, providing different facets of the forecast skill, are simple to calculate and are particularly useful when a forecaster needs to convey the forecast performance to non-technique personnel, such as the general public (Yang et al., 2020). Despite their popularity, these error metrics are often chosen subjectively and the information that each conveys is limited. The distribution-oriented approach, on the other hand, involves analyzing the joint, marginal, and conditional distributions between forecast and observations. It is ideal for forecast analysis as it reveals multifaceted aspects of the forecast quality and can better assist in decision-making when choosing between different modeling schemes or performing cross-

scenario comparisons. The downside of the distribution-oriented approach is the analysis outcome may be too cumbersome to be digested by non-technique personnel. In the next few paragraphs, we discuss the common methods used in each approach in further detail.

When evaluating deterministic forecasts of continuous weather variables, the measure-oriented approach can be further divided into four different categories as shown in *TABLE 2*, with common measures within each category. Distance metrics, usually the smaller the better, summarize the deviation from the forecast to the observations. The key difference between the two popular measures, MAE and RMSE, is that the latter is highly biased for large errors while the former applies the same weight across the errors. The mean bias error (MBE) is the average forecast error representing the systematic error of a model. The MBE, although frequently reported in the meteorological literature, may not be treated as an essential measure. This is because operational forecasts nowadays typically go through a bias correction process before being disseminated to the end-user, and a small MBE is more of a baseline requirement (Yang et al., 2020).

TABLE 2: Classification of measure-oriented approach with common measures within each category

Measure-oriented approach	Common Measures
Distance metrics	• Mean absolute error (MAE) • Root mean square error (RMSE) • Mean bias error (MBE)
Normalized distance metrics	• Normalized mean absolute error (nMAE) • Normalized root mean square error (nRMSE)
Skill scores	• Mean square error skill score (MSESS)
Correlation metrics	• Pearson correlation coefficient (PCC) • Anomaly correlation coefficient (ACC)

Distance metrics are often scale-dependent. Some form of rescaling (or normalization) can be applied, usually by dividing the distance metric by the mean (i.e.,

the mean normalization, see (Wernli et al., 2009)) or the range (i.e., the min-max normalization, see (Hoffman et al., 2018)) of that metric. Nevertheless, none of these normalized metrics (e.g., nMAE and nRMSE) shall be used to compare forecast skills at different locations within different periods. This is because the scaling factor (i.e., the denominator) of such metrics is related to neither the variability nor the uncertainty of the forecasted weather variable, which jointly define the predictability at a given location during a specific period (Yang et al., 2020). In this case, a skill score is recommended for cross-scenario comparison. The skill score is computed by comparing a distance metric (usually the RMSE) of the original forecast to a reference forecast (baseline). The reference forecast could be based on a persistence naïve model, an average of the climatology (typically with a length of 30 years) (Jolliffe & Stephenson, 2012), or a combination of both (Yang et al., 2020). Besides, correlation metrics such as the Pearson correlation coefficient (PCC) are often used to explore the association between the forecast and observations. PCC is mathematically linked to the mean square error skill score (MSESS) (see details in (Murphy, 1988)) and the former is an increasing function of the latter. As weather forecasting has an intrinsic spatial nature, it is necessary to evaluate the forecast skill over a spatial region as a whole. One of the most widely used measures to evaluate spatial forecasts is the anomaly correlation coefficient (ACC). It measures how well the forecast anomalies (i.e., the forecast value less the climatology at a given location) represent the observed anomalies (i.e., the observations less the climatology at the same location). The ACC varies between -1 and 1. In practice, a 60% ACC is usually set as a threshold when determining the skillfulness of synoptic-scale weather forecasts (Zhang et al., 2019). *TABLE 3* provides some representative studies of

weather forecast quality evaluation with a subset of analyzed weather variables and the classical evaluation measures, sorted in chronological order. Among the headers, "Temp.", "Precip.", "Solar", "Wind", "Press.", and "Dist.-oriented" stands for temperature, precipitation, solar irradiance, wind speed, air pressure, and papers utilizing distribution-oriented approach, respectively. Under the measure-oriented approach, the RMSE is the most popular distance metric. Most use cases of the MAE are for evaluating the temperature forecast. Since the majority of these studies are evaluating spatial forecasts, ACC is also prevalently used.

TABLE 3: Representative weather forecast quality evaluation studies with analyzed weather variables and classical evaluation measures, sorted in chronological order

Papers	**Weather variables**					**Measure-oriented**								**Dist.**
	Temp.	Precip.	Solar	Wind	Press.	MAE	RMSE	MBE	nMAE	nRMSE	MSESS	PCC	ACC	**-oriented**
(Kalnay & Dalcher, 1987)					✓								✓	
(Murphy, 1988)	✓	✓					✓				✓	✓		
(Cheng & Steenburgh, 2005)	✓			✓		✓		✓						
(Wernli et al., 2009)		✓					✓	✓		✓				
(Novak et al., 2014)	✓	✓				✓							✓	
(Stern & Davidson, 2015)	✓	✓											✓	
(Vallance et al., 2017)			✓			✓	✓	✓						
(Hoffman et al., 2017)	✓			✓			✓			✓			✓	✓
(Hoffman et al., 2018)							✓	✓		✓	✓		✓	✓
(Zhang et al., 2019)	✓			✓	✓								✓	
(Rasp et al., 2020)	✓					✓	✓						✓	
(Yang et al., 2020)			✓				✓			✓	✓			✓

The measure-oriented approach reduces the information about the forecast quality to a summary measure. In practice, the usage of a single summary measure may not distinguish between forecasts with different behaviors. For instance, Fig. 1 of (Vallance

et al., 2017) shows that when evaluating the performance of solar irradiance forecasts, the RMSE of the two forecasts were the same although they showed very different behaviors. One possible solution to this is to use summary assessment metrics, which are calculated by a weighted sum of multiple normalized measures to avoid shortcomings of individual measures (Hoffman et al., 2018).

On the other hand, the distribution-oriented approach, which provides complete information about the forecast quality, is not limited to certain summary measures. In (Murphy & Winkler, 1987), the joint distribution of the forecast and observations is factored into conditional and marginal distributions in two different ways as follows:

$$p(f,x) = p(x \mid f)p(f)$$
(1) – Calibration-refinement factorization

$$p(f,x) = p(f \mid x)p(x)$$
(2) – Likelihood-base rate factorization

where f and x denote the forecast and observations, respectively. $p(f,x)$ denotes the joint distribution of f and x. The first factorization called calibration-refinement factorization, involves the conditional distribution of the observations given the forecast, $p(x \mid f)$, and the marginal distribution of the forecast, $p(f)$. $p(x \mid f)$ defines the *reliability* of a forecast: if the observations given the forecast are conditionally unbiased, then the forecast is considered completely reliable. $p(f)$ defines the *sharpness* (or variability) of a forecast: if a forecaster always produces the same forecast, then the forecast is said to be not sharp. The second factorization called likelihood-base rate factorization, involves the conditional distribution of the forecast given the observations, $p(f \mid x)$, and the marginal distribution of the observations, $p(x)$. $p(f \mid x)$ defines the *likelihood* of a forecast, while $p(x)$ defines the *uncertainty* or *base rate*. For a perfect

forecast, the $p(f)$ must match the $p(x)$. In-depth details of these terminologies can be found in Chapter 2 of (Jolliffe & Stephenson, 2012) and are summarized in Table 2 of (Murphy, 1993). Common quantification measures of these terminologies are summarized in Table 3 of (Yang et al., 2020). Technical details regarding the link between the measure-oriented approach and the distribution-oriented approach are elucidated in (Murphy, 1993) and (Yang et al., 2020).

The abovementioned distributions can subsequently be visualized using graphical methods for forecast analysis. Histograms, boxplots, and empirical cumulative distribution functions (ECDF) plots can be used to compare the properties of the two marginal distributions ($p(f)$ and $p(x)$). An observation-forecast scatter plot can be used to demonstrate the joint distribution between the observations and forecast. To estimate the two conditional distributions, the kernel conditional density estimation (KCDE) can be used (Yang et al., 2020). Visualization examples of these distributions can be found in Fig 3 - 5 of (Yang et al., 2020).

To summarize, different error measures have their own merits, while they all have some cons. There has been no consensus in the meteorological community on which approach or measure is the best to go with. It is worth noting that the measure-oriented and distribution-oriented approaches are complementary as each has its own use case (Yang et al., 2020). In this research, we use the MAE and observation-forecast scatterplot to investigate the quality of the temperature forecast in Section 4.2, and the prediction performance of the temperature forecast quality in Section 6.3.

Due to the varying predictability of the atmosphere, the weather forecast quality may vary from day to day and region to region. Notable literature predicting the weather

forecast quality is still scarce. (Grönås, 1985) offered one of the first attempts to predict the forecast quality of an ECMWF model. The daily forecast quality (up to 8 days ahead) of surface pressure and 500 hPa geopotential height under T10 resolution was analyzed. Three binary predictors derived from atmospheric flow regimes were tested, while the relationship between these predictors and the forecast quality was found not strong enough. (Kalnay & Dalcher, 1987) predicted the daily forecast quality of 500 mb height and sea level pressure based on the dispersion of the members of an ensemble forecast. Two predictands were tested out, namely, the time when the forecast became unskillful, i.e., when the average ACC dropped below 60%, and the time when the forecast became very good, i.e., when the average ACC reached 80% or above. The proposed method showed promising performance when verified against 14 forecasts over 4 regions. (Palmer & Tibaldi, 1988) uncovered four types of predictors to predict the daily skill of 10 days ahead ECMWF forecasts. The first predictor was related to the consistency of spread between adjacent forecasts, which was influenced by the work in (Kalnay & Dalcher, 1987). The second predictor was based on large-scale atmospheric flow patterns associated with either skillful or unskillful forecasts, which was derived from a regression analysis. The third predictor was the RMSE skill of earlier forecasts, and the last one was the RMSE difference between the forecast predictand (the 500 mb height) and the predictand at the initialization time. (Molteni & Palmer, 1991) tested several variations of the methods and predictors proposed in (Palmer & Tibaldi, 1988) in an operational context. Apart from the abovementioned seminal works, a few more later developments in this area were reviewed in (Kalnay, 2019).

All reviewed works regarding weather forecast quality prediction belong to the field of spatial forecasting. These works leveraged atmospheric state and ensemble weather forecast spread to predict the spatial forecast quality. Most end-users only have access to the deterministic forecasts for a location of interest, while the forecasting spread from ensemble weather forecasts and the atmospheric state data are not accessible. Besides, the temperature forecast we want to utilize is at the surface level, which is more relevant to human activities and their electricity consumption. Hence in this research, we are to build our own wheels to link the quality of deterministic temperature forecasts at multiple weather stations (predictand) with a list of variables that could be related to the predictability of temperature. The variables we have explored are detailed in Section 6.2.3. The prediction outcomes are presented in Section 6.3.

CHAPTER 3: THEORETICAL BACKGROUND

3.1. Multiple Linear Regression

Multiple linear regression (MLR) is a classical statistical technique that models a linear relationship between each independent variable (also known as the predictor or regressor) and the response variable. Compared to the black box approaches such as neural networks and gradient boosting models, MLR is significantly easier to interpret while remaining competitive in providing accurate forecasting results. The model can have the following form:

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

(3) – Linear regression model

where Y is the response variable. $X_1, \dots, X_k$ are the predictors, and $\beta_1, \dots, \beta_k$ are the associated coefficients. β_0 is the intercept term. ε_i are the error terms. The response function is in the following form:

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

(4) – Response function of linear regression model

To construct the confidence intervals and perform statistical inferencing such as the usual F test or Student's t-test, the model needs to stick with the following assumptions:

- The underlying relationship between the dependent variables and the independent variables is linear. Here the dependent variables are assumed numeric values.
- The data is a random sample drawn from the population. The assumption of random sampling implies that (x_i, y_i) and (x_j, y_j) are statistically independent when $i \neq j$.

- The independent variables should not be highly correlated with each other. In other words, the effect of changes in one of the independent variables on the dependent variable does not depend on the values of other independent variables.
- The error terms ε_i are independently and identically normally distributed (IID) with constant variance (homoscedasticity of errors).

The model coefficients can be estimated by minimizing a cost function as follows, which is the sum of squared errors:

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

(5) – Cost function to estimate parameters of linear regression model

where N is the number of observations. This equation has a closed-form solution in matrix form:

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

(6) – Parameter estimation in matrix form

However, solving this equation requires an inverse of $X^T X$, which is nontrivial and can be computationally costly. Methods like Singular Value Decomposition (SVD) or QR Decomposition can be used to calculate the inverse of $X^T X$ (called pseudo-inverse) without actually needing to find an inverse. The model coefficients can also be estimated by gradient descent approaches, which is particularly useful when dealing with a large dataset and the statistical inference is not the primary focus.

There are two types of independent variables: quantitative and qualitative. Quantitative variables are numerical and can be introduced to the model directly. Qualitative variables are sometimes referred to as categorical variables. One way to incorporate categorical variables into regression models is using one-hot encoding. For

instance, in a load forecasting model, the temperature is a quantitative variable, while the hours of a day is a qualitative variable. After applying the one-hot encoding, the coefficients of the hours of a day variable (23 in total) will represent diverse load levels at different hours:

$$\begin{cases} X_1 = 1, \text{if current hour is 0 AM} \\ X_1 = 0, \text{otherwise} \\ X_2 = 1, \text{if current hour is 1 AM} \\ X_2 = 0, \text{otherwise} \\ \qquad\qquad \ldots \\ X_{23} = 1, \text{if current hour is 11 PM} \\ X_{23} = 0, \text{otherwise} \end{cases}$$

(7) – One-hot coding of the hours of a day variable

Interaction effects can be introduced in a regression model when the impact of a predictor on the response variable depends on the values of some other predictors (Hong et al., 2011). In a load forecasting model, the temperature can have an interaction effect with the hours of a day variable. This is because the effect of coincident hour temperature on the load is not independent of different hours within a day.

3.2. Base Model

MLR is one of the earliest and most widely applied techniques in the field of load forecasting (Papalexopoulos & Hesterberg, 1990)(Hong, 2010)(Hong & Fan, 2016). One of the frequently cited MLR models is Tao's Vanilla Benchmark model (hereafter, Vanilla model) (P. Wang et al., 2016)(Hong et al., 2015). Given its accuracy and computational efficiency, it has been used as a benchmark model in recent load forecasting competitions (Hong, Pinson, et al., 2014)(Hong et al., 2016)(Hong et al., 2019). The specification of this model is provided as follows:

$$\widehat{Load_t} = \beta_0 + \beta_1 Trend_t + \beta_2 H_t + \beta_3 W_t + \beta_4 M_t + \beta_5 H_t W_t + \beta_6 T_t + \beta_7 T_t^2 + \beta_8 T_t^3 + \beta_9 T_t H_t + \beta_{10} T_t^2 H_t + \beta_{11} T_t^3 H_t + \beta_{12} T_t M_t + \beta_{13} T_t^2 M_t + \beta_{14} T_t^3 M_t$$

(8) – Tao's Vanilla benchmark model

where $\widehat{Load_t}$ is the coincident response variable forecasted by the predictor variables on the right-hand side of the equation. Among the predictors, $Trend$ is an increasing natural number to model the linear trend of the load within the data history. H_t is a class variable with 24 levels, representing the level shifts of load at 24 hours of a day. W_t is a class variable with 7 levels, representing the level shifts of load at seven days of a week. M_t is a calendar month class variable with 12 levels, representing the level shifts of load at 12 months of a year. T_t is the numerical temperature corresponding to time t. The 3rd order temperature polynomial is used to model the non-linear relationship between the temperature and load. The temperature terms interact with the hours of a day (H_t) and the months of a year (M_t). The hours of a day (H_t) interacts with the days of a week (W_t). Further details about this benchmark model can be found in (Hong et al., 2011).

To achieve better accuracy for a very short forecast horizon, e.g., the day-ahead forecast, we augment the abovementioned model by introducing a lagged load variable. In this research, our forecast horizon is 42 h and we can thus leverage the actual load information 48 hours before the forecast hours. Given the added lagged dependent variable ($Load_{t-48}$), the MLR model turns into a dynamic regression model (DRM). Compared to the Vanilla model, the augmented form leads to one more coefficient to be estimated. This DRM will be used as the base model in this research:

$$\widehat{Load_t} = \beta_0 + \beta_1 Trend_t + \beta_2 H_t + \beta_3 W_t + \beta_4 M_t + \beta_5 T_t + \beta_6 T_t^2 + \beta_7 T_t^3 + \beta_8 H_t W_t + \beta_9 T_t H_t + \beta_{10} T_t^2 H_t + \beta_{11} T_t^3 H_t + \beta_{12} T_t M_t + \beta_{13} T_t^2 M_t + \beta_{14} T_t^3 M_t + \beta_{15} Load_{t-48}$$

(9) – Base model: a dynamic regression model based on the Vanilla model

3.3. Recency Effect

In the context of electric load forecasting, the recency effect refers to the impact of the lagged temperatures (i.e., temperatures of the preceding hours) on the current hour load (Hong, 2010). The family of recency effect models, first introduced in (Hong, 2010), is an extension of the Vanilla model which can dramatically improve the load forecast accuracy. In (P. Wang et al., 2016), the authors went through a model selection process and investigated the number of lagged hourly temperatures and lagged daily moving average temperatures which resulted in the best forecast accuracy. The family of recency effect models can be written as follows:

$$\widehat{Load_t} = \beta_0 + \beta_1 Trend_t + \beta_2 H_t + \beta_3 W_t + \beta_4 M_t + \beta_5 H_t W_t + f(T_t) + \underbrace{\sum_d f(\tilde{T}_{t,d}) + \sum_h f(T_{t-h})}_{\text{recency effect terms}}$$

(10) – Recency effect model based on the Vanilla model

where T_{t-h} is the lagged hourly temperature at the preceding h^{th} hour ($h = 0, 1, 2, ...$). $\tilde{T}_{t,d}$ is the daily moving average temperature of the preceding d^{th} day:

$$\tilde{T}_{t,d} = \frac{1}{24} \sum_{h=24d-23}^{24d} T_{t-h}, \quad d = 0, 1, 2, ...$$

(11) – Lagged moving average temperature

and $f(T_t)$ denotes the regression terms associated with T_t with interaction effects included:

$$f(T_t) = \beta_6 T_t + \beta_7 T_t^2 + \beta_8 T_t^3 + \beta_9 T_t H_t + \beta_{10} T_t^2 H_t + \beta_{11} T_t^3 H_t + \beta_{12} T_t M_t + \beta_{13} T_t^2 M_t + \beta_{14} T_t^3 M_t$$

(12) – Regression terms related to temperature

Following (P. Wang et al., 2016), this framework can be used to investigate the forecasting performance of models with diverse levels of complexity. In this research, the values of h and d are enumerated from 0 to 24, and 0 to 3, respectively. In other words, we are exploring 25 * 4 = 100 individual models with different levels of complexity.

It is important to understand that this framework follows an incremental manner when introducing the preceding hour temperatures (notice the two summation signs in equation (10)). For instance, letting $h = 2$ will add the regression terms associated with T_{t-2} as well as T_{t-1} to the equation. This is rational as the current hour load should be more relevant to the temperature information of the current hour and the more recent preceding hours. When adding the temperatures that are further in the past, the more recent hour temperatures should be kept within the model. Nevertheless, in the ex-ante forecasting settings, such a design lacks the flexibility to drop the recent hour temperatures with significant forecast errors while keeping the ones with better quality in the further past hours. We will discuss this limitation further in Chapter 7 along with our case study and suggest future research directions in Chapter 8.

3.4. Sliding Simulation

Day-ahead load forecasting is an essential element in power system planning, operations, energy trading, and so forth. Normally, the forecast will be provided by the load forecasters in the morning, e.g., 7am, for the 24 hours of the next day (from 1am to 12am). To evaluate the out-of-sample performance of the models, we use a sliding simulation technique that imitates the day-ahead load forecasting procedure.

For the 9 years (2010-2018) data history (further introduced in Section 4.1), the most recent four years of the load data (2015-2018) is held out for the out-of-sample test. *FIGURE 4* shows the sliding simulation technique used to generate day-ahead hourly load forecasts in the test period. The training data is with a fixed length of around 2 years (2 * 365 – 1 days plus 6 hours). It consists of load and actual temperature being used for parameter estimation. The forecast origin (i.e., the last data point in the training data) is at 6am on the current day. The forecast horizon (i.e., the length of forecast into the future) is 42 hours, which is from 7am of the current day to the last hour of the next day (ending at 12am). For day-ahead load forecasting, we only evaluate the hourly forecast of the next day (from 1am to 12am). After generating the forecast for a day, we advance the forecast origin by 24 hours to forecast the next 42 hours in the forecast horizon. We repeat this process until we generate the forecasts for all four years from 2015 to 2018.

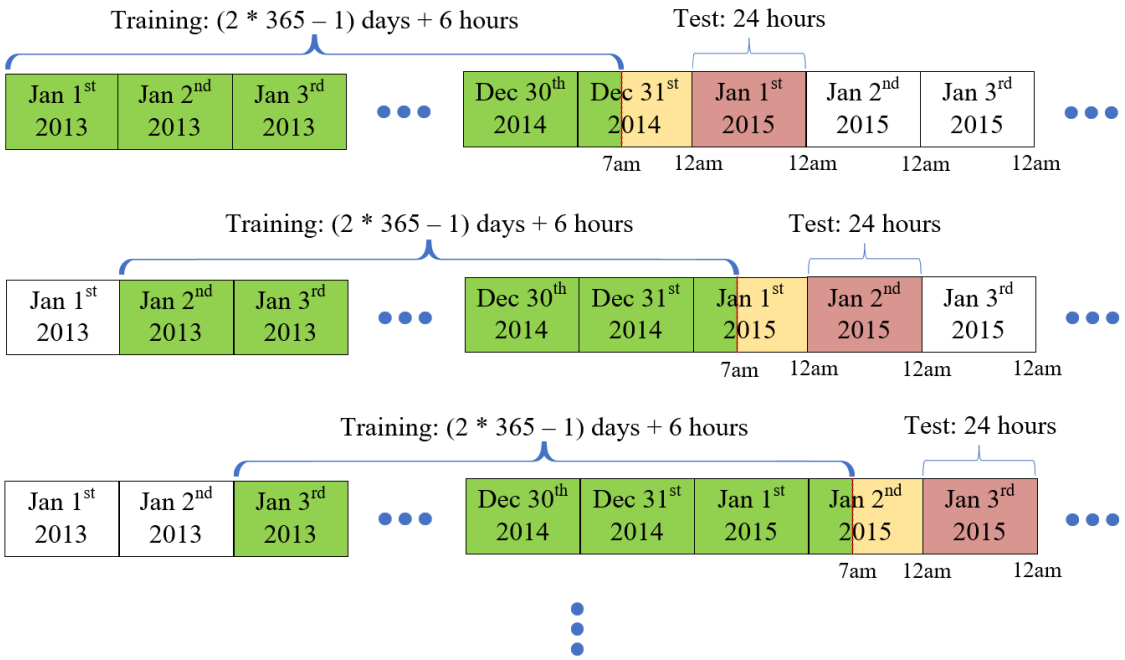

FIGURE 4: Sliding simulation for day-ahead load forecasting

3.5. Forecast Evaluation

Ex-post and ex-ante forecasts are both evaluated in this research. In the forecast period, ex-post forecasts are generated using the hourly actual temperature as input. As the state-of-the-art literature mostly uses ex-post forecast for model selection, we include it as a benchmark method in this research and discuss it further in Chapter 5. On the other hand, the ex-ante forecasts are generated using the hourly day-ahead temperature forecast. It is worth mentioning that we don't use temperature forecast for model fitting, as doing so may result in biased parameter estimation due to noise in the temperature forecast.

An error measure should be chosen when evaluating the performance of a forecasting model. Different types of error measures are mentioned in the load forecasting literature, each with its own applications. The mean absolute percentage error (MAPE) is one of the most widely used error measures in the load forecasting field, due to its scale-independency and interpretability. However, a major drawback of using MAPE is the error can be infinite or undefined for zero or close-to-zero actual values. In this research, since no loads are close to zero, we use MAPE as the error measure:

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - \widehat{y_i}|}{y_i}$$

(13) – Error measure: MAPE

where y_i and $\widehat{y_i}$ are the actual values and predicted values, and N is the number of observations in the forecast period.

3.6. Weather Station Selection

The quality of temperature input to a load forecasting model is of critical importance for better load forecast accuracy. As we have temperature information from 18 weather stations, weather station selection (WSS) is necessary to refine the temperature input to the load forecasting model.

In this research, we follow the algorithm proposed in (Hong et al., 2015) for WSS during the test years. The reason why we follow this approach is twofold. First, (Hong et al., 2015) was the first paper published on solving the WSS problem specifically. Second, the efficiency of the algorithm is trackable: it has frequently been adopted in the load forecasting literature (e.g., (Xie & Hong, 2016), (Sobhani et al., 2019), and (Neto & Hippert, 2020)) and due to its transparency and simplicity, the method is currently used by many power companies (Y. Wang et al., 2019).

In our case study, we use the Vanilla model introduced in Section 3.2 as the load forecasting model for WSS in the recent six years (2013-2018). Following (Hong et al., 2015), the training data is comprised of historical load and temperature data. We first use a window of three years (2010-2012) historical data to determine the weather station selection for 2013. We then advance the window by one year and use the data from 2011 to 2013 to determine the weather station selection for 2014. This process is repeated until we have the selected weather stations for all six years. Thereafter, a composite temperature series is obtained by a simple averaging of the selected stations at each year.

CHAPTER 4: DATA

4.1. Load and Weather Data

Our case study is based on data from a medium-sized power utility on the east coast of the US. The load time series consists of 9 years (2010-2018) of hourly load data from three supply areas within the utility's service territory. The three supply areas are adjacent to each other and named SA1, SA2, and SA3, respectively.

Weather data collected at 18 weather stations were purchased by the power utility from a commercial source. The stations are located within or near its service territory. The weather data consists of 9 years (2010-2018) of historical hourly temperature data and 6 years (2013-2018) of day-ahead hourly temperature forecast data. The temperature forecast data are released at 7am each day and forecast the hours throughout the next day. In this research, we only use the temperature forecast during the next day (24 h, from 1am to 12am) to generate the ex-ante load forecast.

4.2. Exploratory Data Analysis

The composite historical and forecasted temperature series are both obtained by averaging the stations selected in Section 3.6. *FIGURE 5* shows six years (2013-2018) of load and composite historical temperature time series under SA1. Strong seasonality can be observed in both series. *FIGURE 6* shows the load series under SA2. The load of 3 days in 2016 (October 8^{th} - 10^{th}) and 5 days in 2018 (September 13^{th} - 17^{th}) are extraordinarily lower than usual. These could be caused by circuit-level outages. Hence, we excluded the 8 days from our error analysis. *TABLE 1* provides the statistics of load data under each supply area. Based on the average load, SA2 is a larger supply area, while SA1 is the smallest.

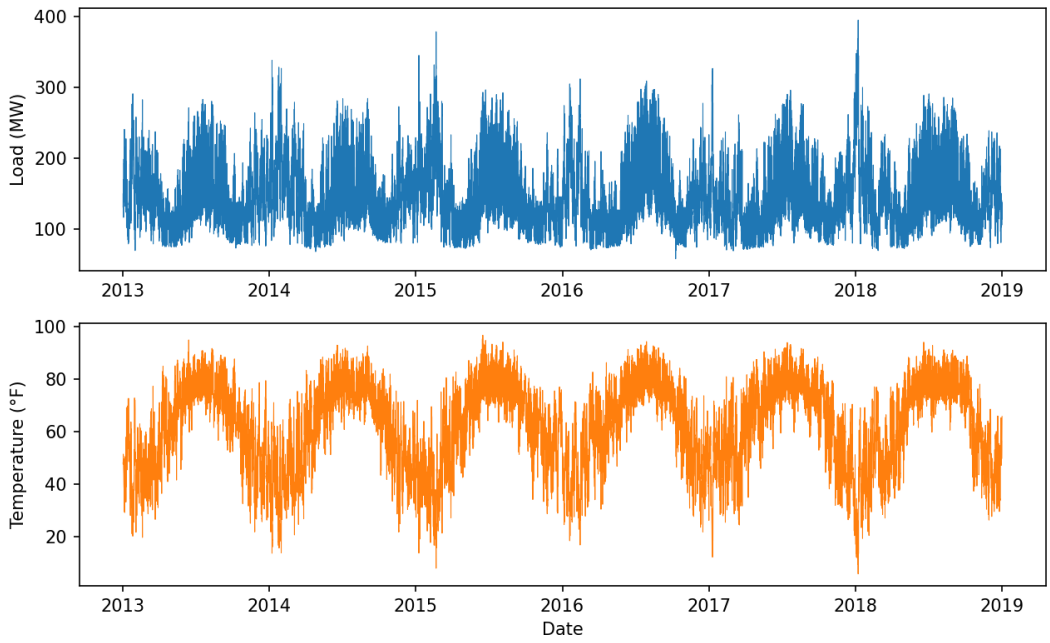

FIGURE 5: Load and historical temperature time series under SA1 (2013 – 2018)

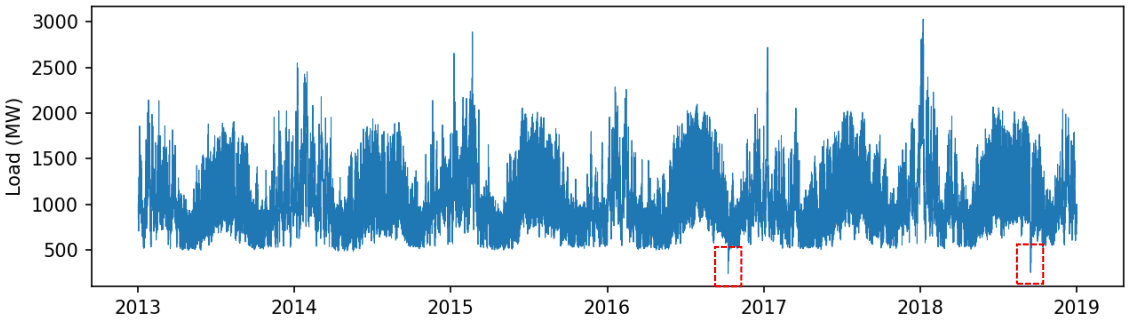

FIGURE 6: Load time series under SA2 (2013 – 2018)

TABLE 4: Statistics of Load Data (in MW)

Supply Area	Days Excluded in Error Analysis	Mean	Std.	Min.	Max.
SA1	0	149.3	48.3	16.1	394.5
SA2	8	1042.9	350.2	240.1	3027.8
SA3	0	230.1	81.9	99.8	553.1

FIGURE 7 shows the load vs. historical temperature plot under SA1. We observe a strong nonlinear correlation (the typical "hockey stick" shape) between the two

variables. On the left arm, when the temperature drops below 60 °F and keeps decreasing, the load demand rises due to the incremental heating demand during winter. On the right arm, the load rises as the temperature climbs up, due to the incremental cooling needs during summer.

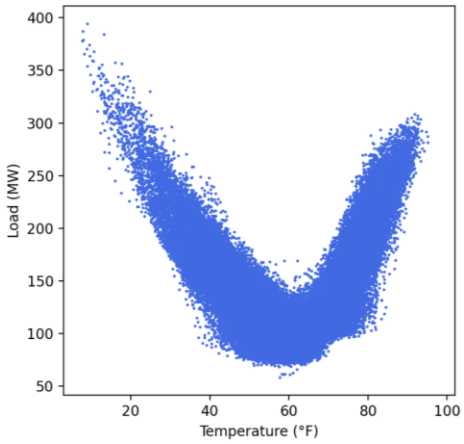

FIGURE 7: Load - historical temperature scatterplot, SA1 (2013 – 2018)

TABLE 5 gives the statistics of historical temperature readings at each weather station. Both the average temperature and the standard deviations vary at each station.

FIGURE 8 demonstrates the variance of the actual hourly temperature during a winter day in 2015. The recorded temperatures among the weather stations are having greater variations during midnight and early morning.

TABLE 5: Statistics of actual temperature reported at each weather station

Station	Temperature (°F)	
	Mean	Std.
1	61.5	17.6
2	59.1	17.2
3	56.6	16.6
4	62.0	16.8
5	63.5	16.0
6	62.1	16.6
7	63.2	16.3
8	63.3	16.4
9	52.8	16.2
10	62.7	16.9
11	59.9	17.2
12	59.7	16.9
13	58.7	17.0
14	64.3	14.6
15	60.5	17.2
16	64.2	15.9
17	59.1	17.5
18	64.2	15.1
Range	52.8 - 64.3	14.6 - 17.6

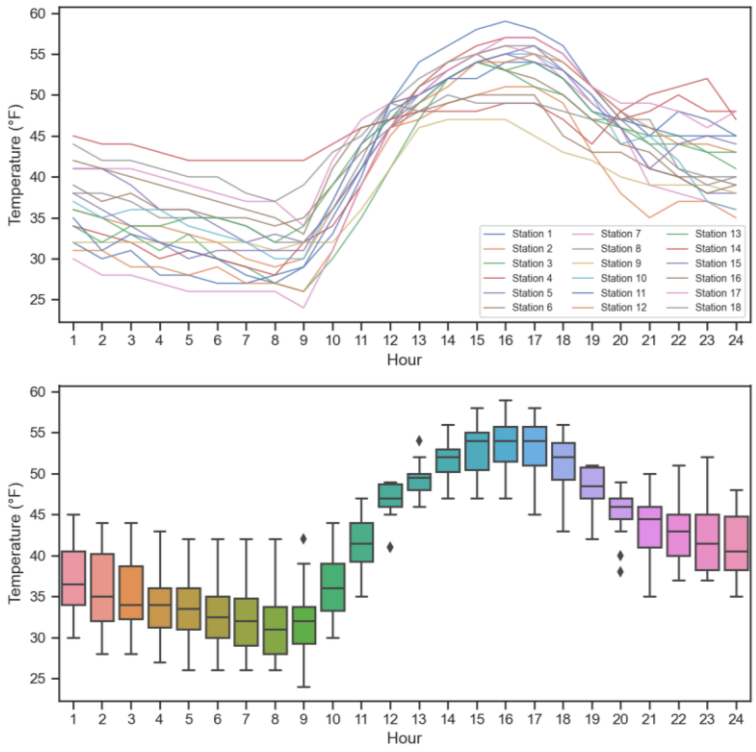

FIGURE 8: Line plot (up) and boxplots (down) of historical temperature reported at each weather station, 1/25/2015

Based on weather station selection, the composite day-ahead temperature forecast is used for ex-ante load forecasting. There are a few (less than 1%) missing values in the day-ahead temperature forecast data. The 2 days ahead temperature forecast is used to fill in these missing values. When the 2 days ahead forecast is not available, the actual temperature is used. *FIGURE* 9 visualizes the historical (actual) and forecasted temperature during the week of 1/25/2015. Overall, the forecast well captures the major trend of the actual temperature. Nevertheless, the timing and magnitude of daily peaks (such as 1/27/2015) and troughs (such as 1/29/2015) were not captured precisely. Meanwhile, the forecasted temperature can either be above or below the historical temperature for prolonged periods, rather than showing as a random noise around the historical temperature.

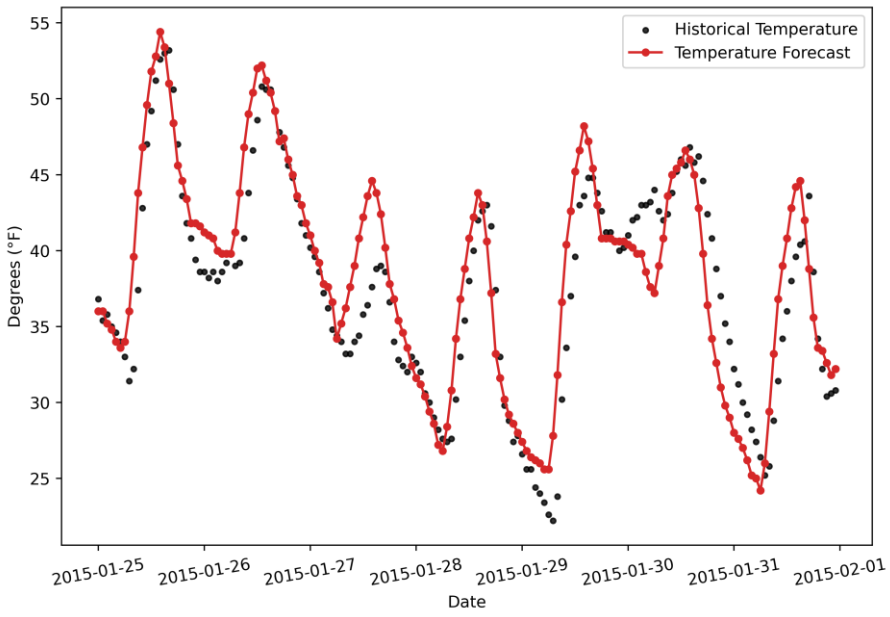

FIGURE 9: Historical (black dots) and forecasted (red) temperature under SA1 (1/25/2015-1/31/2015)

FIGURE 10 shows the joint and marginal distribution of the day-ahead temperature forecast and actual temperature. Overall, the forecasted temperature is well

aligned with the actual temperature with a few salient over forecasts when the actual temperature is around 45 °F and a few salient under forecasts when the actual temperature is within 60 to 80 °F. Taking a closer look at the density contours, the forecasted temperature drifts slightly below the identity line when the actual temperature is below 80 °F, and slightly above the identity line when the actual temperature is above 80 °F. *FIGURE 11* shows a histogram of temperature forecast errors for SA1. The errors are calculated by using historical temperatures minus the forecast. The distribution is close to normal with a slight skew to the left. Some extreme negative values are observed (due to over forecast). *FIGURE 12* shows a time series plot of the absolute temperature forecast errors. Due to the lack of either the day-ahead or 2 days ahead temperature forecast data, the errors from June 18, 2014, to June 28, 2014, are shown as 0. Some extreme errors are observed during 2014 and 2015 and were kept within our analysis. *TABLE 6* shows a heatmap of statistics on the absolute errors of composite temperature forecasts. The greener color indicates the years with better accuracy on average, lower maximum error, and a smaller standard deviation of the absolute errors. All three statistics show a drastic improvement beyond 2015, with a minimized error and error standard deviation during 2018. Besides, the mean absolute error shows a sustained improvement ever since 2013 at most times. Similar findings were observed under the composite temperature forecasts of SA2 and SA3. To avoid verbose representation, we did not show them in this section.

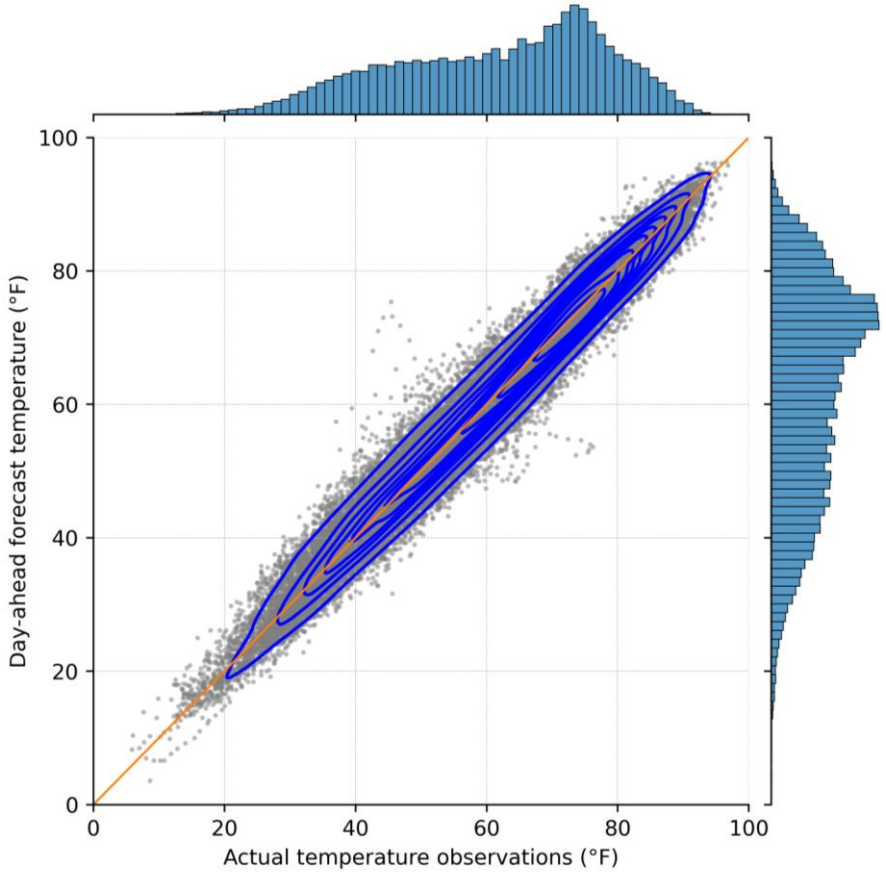

FIGURE 10: Joint and marginal distribution of day-ahead temperature forecast and observations under SA1 (2013-2018) with contour lines in blue showing the kernel densities

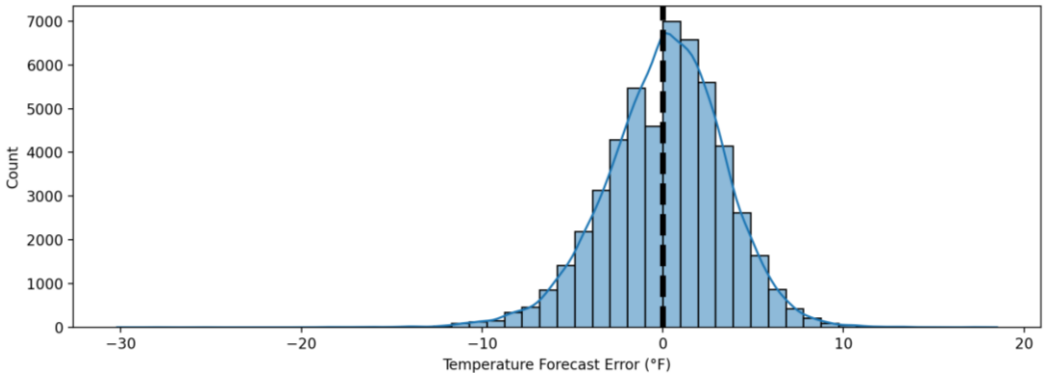

FIGURE 11: Histogram of temperature forecast error, composite temperature forecasts under SA1 (2013-2018)

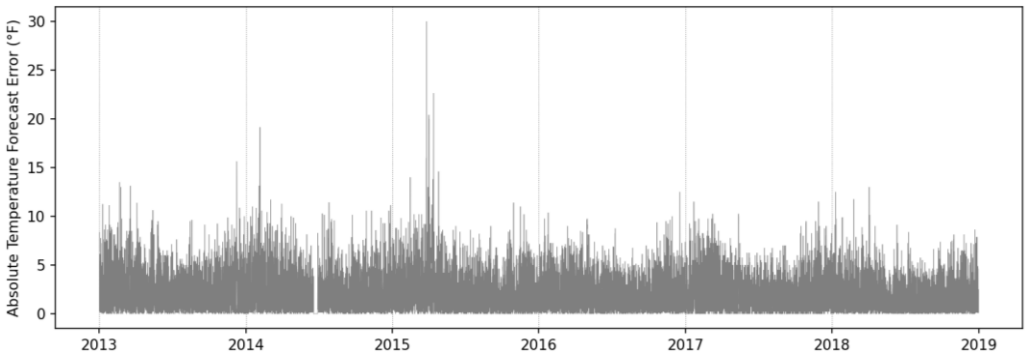

FIGURE 12: Time series plot of absolute temperature forecast error, composite temperature forecasts under SA1

TABLE 6: Heatmap of mean absolute errors, maximum absolute errors, and standard deviation of absolute errors for each year, composite temperature forecasts under SA1

Year	Mean AE (°F)	Max AE (°F)	STD. of AE (°F)
2013	2.48	15.6	1.94
2014	2.44	19.1	1.99
2015	2.42	30.0	2.25
2016	2.09	12.5	1.65
2017	2.17	11.5	1.72
2018	1.90	13.0	1.58

CHAPTER 5: MODEL SELECTION BASED ON EX-ANTE FORECAST

Load forecasting models were commonly developed based on the historical ex-post forecast accuracy. In this research, we denote the approach of load forecasting model selection based on ex-post forecast accuracy as the benchmark framework "M0". In this chapter, we propose using the ex-ante forecast accuracy to select load forecasting models (denoted framework "M1"). In the following sections, we introduce the workflow of our methodology, the forecasting models, and compare the out-of-sample forecasting performance between M1 and M0.

5.1. Methodology

5.1.1 Overall Procedure

FIGURE 13 shows a high-level workflow of the overall methodology. The raw data (load and temperature) will first go through the data cleansing and weather station selection process. Next, the available data history will be partitioned into training data (years 2013 and 2014), validation data (years 2015 and 2016), and out-of-sample test data (years 2017 and 2018). The training data is used for model fitting and parameter estimation of each candidate model. The validation data is used for model performance evaluation and selection by each framework (M1 or M0). In the end, the selected model by either method will refit the validation data and produce the forecast in the test period.

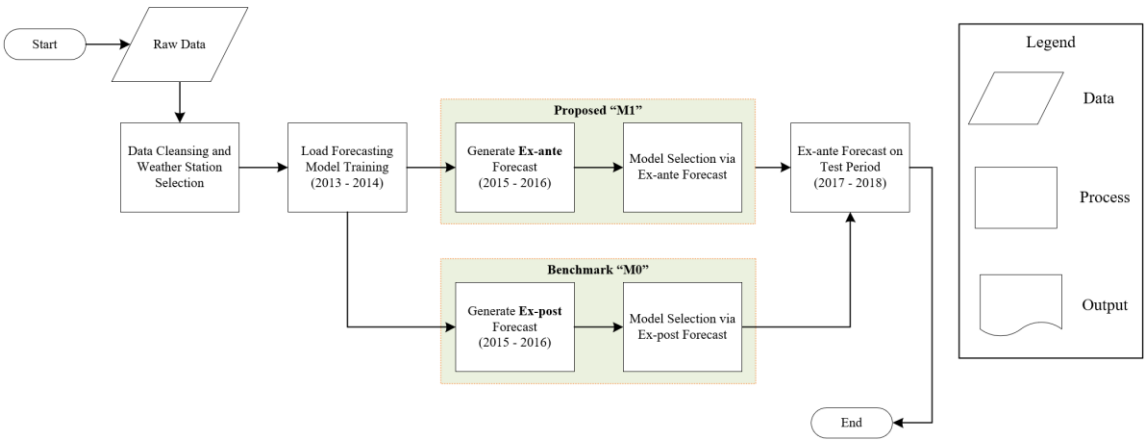

FIGURE 13: High-level workflow of model selection based on ex-ante or ex-post forecast accuracy

5.1.2 Forecasting Models

The candidate models are from a family of recency effect models, as introduced in Section 3.3. Following (P. Wang et al., 2016), we model the recency effect using a trial-and-error method, by varying the value of d (defining the number of lagged moving average temperatures) from 0 to 3, and the value of h (defining the number of lagged hourly temperatures) from 0 to 24. We explore a total of 100 (d, h) pairs.

5.2. Results and Discussion

5.2.1 Model Selection Step

We first examine the load forecasting performance on the validation data (years 2015 and 2016). *TABLE 7* shows a heatmap of the ex-post forecasting MAPE values. A greener color indicates a lower MAPE value, while a redder color indicates a higher one. The bold values indicate the best (d, h) pairs for each supply area. The best model is identified as (1, 6), (1, 10), and (1, 11) for SA1, SA2, and SA3, respectively. *TABLE 8* shows a heatmap of the ex-ante forecasting MAPE values. The best model is identified as (1, 1), (1, 3), and (1, 1) for SA1, SA2, and SA3, respectively.

TABLE 7: Heatmap of the ex-post forecasting MAPE values (in %) for recency effect modeling on the validation data (years 2015 and 2016)

SA1				
$h \setminus d$	0	1	2	3
0	5.124	4.414	4.477	4.576
1	4.678	4.248	4.339	4.430
2	4.463	4.164	4.268	4.358
3	4.340	4.126	4.235	4.326
4	4.271	4.106	4.219	4.307
5	4.234	4.097	4.214	4.299
6	4.207	**4.095**	4.215	4.295
7	4.192	4.099	4.221	4.298
8	4.180	4.105	4.225	4.302
9	4.175	4.108	4.226	4.302
10	4.171	4.113	4.230	4.305
11	4.164	4.114	4.228	4.304
12	4.158	4.116	4.230	4.306
13	4.157	4.123	4.238	4.312
14	4.163	4.132	4.246	4.322
15	4.168	4.139	4.253	4.332
16	4.176	4.147	4.260	4.342
17	4.188	4.154	4.269	4.355
18	4.202	4.163	4.276	4.365
19	4.217	4.176	4.285	4.377
20	4.233	4.187	4.294	4.389
21	4.250	4.203	4.308	4.403
22	4.274	4.221	4.324	4.418
23	4.295	4.235	4.336	4.434
24	4.309	4.245	4.346	4.444

SA2				
$h \setminus d$	0	1	2	3
0	4.681	3.771	3.791	3.873
1	4.132	3.512	3.553	3.635
2	3.867	3.395	3.444	3.522
3	3.729	3.343	3.393	3.471
4	3.640	3.317	3.367	3.443
5	3.578	3.308	3.355	3.429
6	3.528	3.301	3.346	3.420
7	3.498	3.303	3.345	3.416
8	3.472	3.297	3.344	3.414
9	3.450	3.296	3.346	3.415
10	3.428	**3.294**	3.346	3.412
11	3.413	3.295	3.347	3.413
12	3.405	3.301	3.353	3.418
13	3.398	3.307	3.359	3.425
14	3.395	3.311	3.365	3.431
15	3.388	3.311	3.367	3.435
16	3.385	3.316	3.372	3.439
17	3.388	3.317	3.374	3.443
18	3.397	3.321	3.379	3.452
19	3.407	3.326	3.385	3.462
20	3.419	3.332	3.394	3.472
21	3.432	3.343	3.408	3.487
22	3.445	3.355	3.423	3.504
23	3.457	3.366	3.438	3.520
24	3.468	3.375	3.451	3.535

SA3				
$h \setminus d$	0	1	2	3
0	5.053	4.205	4.234	4.315
1	4.596	3.985	4.024	4.110
2	4.352	3.887	3.932	4.017
3	4.227	3.839	3.889	3.971
4	4.148	3.809	3.864	3.946
5	4.086	3.793	3.846	3.925
6	4.036	3.786	3.839	3.915
7	3.994	3.783	3.836	3.910
8	3.957	3.776	3.826	3.900
9	3.922	3.763	3.811	3.883
10	3.902	3.755	3.800	3.874
11	3.887	**3.752**	3.797	3.875
12	3.877	3.753	3.798	3.879
13	3.867	3.756	3.801	3.880
14	3.868	3.762	3.807	3.884
15	3.872	3.773	3.823	3.897
16	3.872	3.781	3.834	3.907
17	3.873	3.789	3.845	3.919
18	3.878	3.798	3.859	3.931
19	3.885	3.806	3.873	3.947
20	3.900	3.822	3.889	3.963
21	3.914	3.837	3.906	3.980
22	3.930	3.850	3.917	3.996
23	3.944	3.865	3.930	4.015
24	3.960	3.873	3.939	4.026

TABLE 8: Heatmap of the ex-ante forecasting MAPE values (in %) for recency effect modeling on the validation data (years 2015 and 2016)

SA1				
$h \setminus d$	0	1	2	3
0	7.170	6.081	6.104	6.165
1	6.852	**5.983**	6.034	6.085
2	6.678	5.995	6.056	6.105
3	6.534	6.010	6.078	6.120
4	6.445	6.043	6.115	6.151
5	6.403	6.094	6.168	6.202
6	6.380	6.147	6.232	6.262
7	6.375	6.218	6.311	6.337
8	6.379	6.293	6.396	6.417
9	6.383	6.363	6.474	6.489
10	6.398	6.438	6.552	6.558
11	6.413	6.500	6.623	6.618
12	6.430	6.550	6.677	6.660
13	6.457	6.591	6.716	6.690
14	6.494	6.639	6.762	6.731
15	6.509	6.672	6.794	6.761
16	6.523	6.688	6.808	6.777
17	6.542	6.690	6.809	6.777
18	6.566	6.689	6.798	6.770
19	6.586	6.689	6.793	6.768
20	6.600	6.690	6.798	6.781
21	6.610	6.697	6.809	6.795
22	6.636	6.716	6.828	6.813
23	6.668	6.750	6.861	6.848
24	6.715	6.785	6.895	6.881

SA2				
$h \setminus d$	0	1	2	3
0	7.021	5.813	5.818	5.887
1	6.679	5.738	5.768	5.837
2	6.473	5.745	5.790	5.855
3	6.291	**5.737**	5.790	5.852
4	6.184	5.751	5.808	5.864
5	6.156	5.795	5.853	5.898
6	6.145	5.839	5.900	5.936
7	6.122	5.883	5.947	5.973
8	6.098	5.931	5.995	6.012
9	6.077	5.970	6.038	6.044
10	6.063	6.003	6.076	6.072
11	6.034	6.012	6.088	6.072
12	6.010	6.013	6.094	6.069
13	6.006	6.034	6.112	6.077
14	5.985	6.056	6.125	6.089
15	5.958	6.069	6.134	6.096
16	5.944	6.073	6.135	6.097
17	5.945	6.072	6.128	6.092
18	5.955	6.083	6.132	6.097
19	5.974	6.094	6.139	6.112
20	6.005	6.115	6.150	6.133
21	6.038	6.142	6.175	6.160
22	6.081	6.178	6.208	6.197
23	6.124	6.224	6.247	6.235
24	6.163	6.255	6.277	6.264

SA3				
$h \setminus d$	0	1	2	3
0	7.127	5.965	5.982	6.036
1	6.876	**5.909**	5.932	5.988
2	6.720	5.936	5.963	6.017
3	6.553	5.939	5.971	6.023
4	6.425	5.941	5.970	6.018
5	6.344	5.949	5.980	6.021
6	6.289	5.968	6.004	6.036
7	6.239	5.997	6.037	6.061
8	6.182	6.026	6.071	6.086
9	6.128	6.043	6.088	6.096
10	6.096	6.072	6.115	6.115
11	6.074	6.083	6.125	6.119
12	6.072	6.078	6.120	6.111
13	6.072	6.077	6.121	6.110
14	6.072	6.100	6.139	6.131
15	6.063	6.127	6.166	6.163
16	6.061	6.148	6.186	6.190
17	6.068	6.159	6.201	6.210
18	6.080	6.168	6.217	6.226
19	6.100	6.174	6.226	6.238
20	6.125	6.186	6.240	6.256
21	6.154	6.215	6.272	6.289
22	6.189	6.259	6.319	6.334
23	6.219	6.293	6.354	6.364
24	6.251	6.311	6.381	6.392

The only difference between the two types of forecast is whether the actual or forecasted temperature is used. Comparing the two aforementioned tables, our findings are threefold. First, the ex-ante forecasts have higher MAPEs than the ex-post counterparts. This is because when generating the ex-ante load forecast using temperature forecast, the associated temperature forecast errors introduce additional errors on top of the existing ex-post load forecast errors. The contribution of temperature forecast error to the ex-ante load forecast error can be quantified by comparing the MAPE values of the two tables under the same recency effect model. Second, the accuracy of both ex-post and ex-ante forecasts starts to degrade when introducing more recency effect terms after a certain point. Third, compared to M0 (model selection based on the ex-post forecast), the M1 framework (model selection based on the ex-ante forecast) picks the models with the same value of $d = 1$, but significantly smaller values of h (more parsimonious models with fewer lagged hourly temperatures).

For instance, under SA1, the lowest MAPE based on the ex-post and ex-ante forecast is achieved at (1, 6) and (1, 1), respectively. When $d = 1$, the ex-post forecast accuracy keeps improving when varying h from 0 to 6. This is because the additional lagged hourly temperatures provide extra signals to assist model the load variations. The ex-post forecast accuracy starts to degrade when additional lagged hourly temperatures (i.e., a larger h value) were introduced on top of the model (1, 6). This is due to overparameterization. On the other hand, the ex-ante forecast accuracy only improves when increasing h from 0 to 1. The accuracy starts to degrade when additional lagged hourly temperatures were included. This can be explained by the fact that, in the ex-ante forecasting settings, the additional lagged temperatures also bring in temperature forecast

errors from the past hours. In other words, the ex-ante forecast accuracy degrades when the extra signal provided by the lagged temperatures is too marginal to overcompensate the detrimental impact of temperature forecast errors. We denote this finding as the "signal-error trade-off".

5.2.2 Out-of-sample Test

Based on the performance of validation data, *TABLE 9* shows the MAPE values on the test data (years 2017 and 2018). A table cell with red borders indicates the model selected based on the ex-post forecast (M0, referring to *TABLE 7*). A table cell with blue borders indicates the model selected based on the ex-ante forecast (M1, referring to *TABLE 8*). The bold values indicate the ground truth of the best model on the test data.

Compared to the ex-post forecasting MAPE values shown in *TABLE 7*, the MAPE values in *TABLE 8* provide a closer estimate of the MAPE values on the test data (*TABLE 9*). This provides evidence that the ex-ante forecast on the validation data better reflects the genuine performance of a load forecasting model on the unseen test set. Besides, for SA1, the model selected by M1 (the ex-ante forecast) outperforms the M0 counterpart by around 1.8% (from 5.766% to 5.660%). For SA2, the ex-ante forecast selects the exact best model on the test data. Compared to the M0 counterpart, the relative MAPE reduction by using the M1 framework is around 6.3% (from 5.844% to 5.476%). For SA3, the model selected by M1 results in a minor increase in MAPE (from 5.735% to 5.781%) compared to M0. This can be explained by the fact that the model selected by M1 may underperform when the temperature forecast quality gets worse in the test period (detailed in Section 7.2.2, *TABLE 19*). In sum, the model selected by the ex-ante forecast leads to superior accuracy in most cases, with an average improvement of 2.4%, without

introducing additional computational costs. This provides evidence that the ex-ante forecast performance should be focused on in the model selection stage.

TABLE 9: Heatmap of the ex-ante forecasting MAPE values (in %) for recency effect modeling on the test data (years 2017 and 2018)

SA1				
$h \setminus d$	0	1	2	3
0	7.065	5.850	5.843	5.939
1	6.650	5.660	5.675	5.770
2	6.380	5.612	5.636	5.729
3	6.191	**5.604**	5.639	5.729
4	6.104	5.643	5.685	5.771
5	6.073	5.705	5.755	5.835
6	6.056	5.766	5.826	5.902
7	6.041	5.819	5.887	5.963
8	6.025	5.867	5.945	6.021
9	6.024	5.921	6.000	6.079
10	6.045	5.982	6.066	6.145
11	6.086	6.048	6.134	6.206
12	6.134	6.098	6.182	6.252
13	6.165	6.139	6.224	6.286
14	6.183	6.174	6.263	6.320
15	6.211	6.224	6.315	6.372
16	6.255	6.276	6.376	6.432
17	6.316	6.327	6.432	6.493
18	6.372	6.381	6.483	6.549
19	6.429	6.427	6.531	6.602
20	6.478	6.471	6.582	6.659
21	6.521	6.507	6.626	6.709
22	6.568	6.541	6.671	6.755
23	6.611	6.575	6.712	6.792
24	6.645	6.602	6.744	6.818

SA2				
$h \setminus d$	0	1	2	3
0	6.914	5.650	5.772	5.882
1	6.508	5.509	5.641	5.752
2	6.220	5.486	5.621	5.726
3	6.019	**5.476**	5.617	5.713
4	5.918	5.490	5.633	5.725
5	5.882	5.539	5.685	5.772
6	5.871	5.586	5.734	5.815
7	5.859	5.643	5.793	5.867
8	5.848	5.696	5.846	5.915
9	5.844	5.764	5.919	5.985
10	5.864	5.844	6.006	6.070
11	5.877	5.884	6.053	6.115
12	5.924	5.931	6.107	6.162
13	5.950	5.979	6.154	6.201
14	5.942	5.999	6.175	6.214
15	5.938	6.021	6.190	6.227
16	5.949	6.041	6.201	6.236
17	5.966	6.042	6.201	6.235
18	5.992	6.057	6.218	6.249
19	6.027	6.079	6.234	6.268
20	6.049	6.096	6.243	6.280
21	6.066	6.098	6.243	6.283
22	6.081	6.110	6.263	6.305
23	6.100	6.136	6.296	6.337
24	6.110	6.151	6.318	6.359

SA3				
$h \setminus d$	0	1	2	3
0	7.167	5.915	5.984	6.072
1	6.830	5.781	5.858	5.962
2	6.593	5.760	5.834	5.938
3	6.408	5.746	5.822	5.920
4	6.265	5.736	5.816	5.909
5	6.160	5.726	5.811	5.900
6	6.084	5.722	5.813	5.896
7	6.016	5.725	5.816	5.895
8	5.945	5.728	5.819	5.893
9	5.886	5.733	5.825	5.898
10	5.848	5.735	5.835	5.910
11	5.821	5.735	5.838	5.915
12	5.806	5.728	5.836	5.915
13	5.783	5.717	5.825	5.908
14	5.746	5.705	5.814	5.902
15	5.706	5.692	5.799	5.892
16	5.690	5.680	5.788	5.886
17	5.685	5.667	5.777	5.880
18	5.680	**5.655**	5.768	5.874
19	5.685	5.666	5.776	5.883
20	5.701	5.693	5.797	5.904
21	5.719	5.719	5.820	5.924
22	5.741	5.744	5.844	5.945
23	5.762	5.767	5.871	5.969
24	5.783	5.774	5.881	5.977

CHAPTER 6: TEMPERATURE FORECAST QUALITY PREDICTION

Empirical case studies reviewed in Section 2.2 (e.g., (Segarra et al., 2019), (Methaprayoon et al., 2007), and (Chitalia et al., 2020)) have shown that the ex-ante load forecast accuracy can be impacted by the levels of weather forecast accuracy. One way to make better model selection decisions is to understand the quality of future weather forecasts and select different load forecasting models at diverse levels of weather forecast accuracy. This requires predicting the accuracy of future weather forecasts given the historical weather forecast information. Since temperature is our forecasted weather variable in day-ahead load forecasting, we experiment with several approaches, including a few baseline models, to predict the quality of the day-ahead temperature forecasts in this chapter.

In the following sections, we first introduce some background information about the forecasting problem. Since the prediction of day-ahead temperature forecast accuracy will eventually be used for load forecasting, we focus on modeling the accuracy of composite temperature forecast for a subset of hours to create a more impactful analysis. Thereafter, we demonstrate our methodology to tackle this forecasting problem and present the experiment results at the end.

6.1. Background

6.1.1 The Forecasting Problem

In the load forecasting problem, the temperature variable is based on a combination of weather stations (refer to Section 3.6). Hence, it is required to predict the quality of the composite temperature forecasts. There are two ways to approach this problem. One is to predict the quality of the temperature forecasts at each weather station and combine these

predictions based on the results of weather station selection. The other one is to predict the quality of the composite temperature forecast directly.

We followed the first approach at the beginning while noticing several major challenges. First, the stochastic nature of weather can lead to diverse forecasting biases at different weather station locations. In this case, it is extremely challenging to design a single method that can generate robust predictions at all locations. Second, the temperature forecasts at the individual station level are prone to data quality issues (e.g., data errors, missing values, etc.). This makes the historical training data even noisier and eventually leads to failures of most machine learning algorithms. Third, our goal is to understand the quality of the composite temperature forecasts for load forecasting. Even though we had a reliable prediction of absolute temperature forecast errors at each station level (refer to Section 6.2.1, our prediction does not provide the signs of errors, i.e., positive or negative), the task of combining the non-negative predicted errors is nontrivial due to the compensation of positive and negative predicted errors.

On the other hand, by averaging temperatures from multiple weather stations, the second approach helps to alleviate the forecasting biases and data errors at individual weather stations. This creates an easier task for the machine learning algorithms to learn from historical data. Besides, the output of the second approach can directly be leveraged by the load forecasting procedure. Hence, we proceed with the second approach in this chapter, which is to model the quality of composite temperature forecasts.

6.1.2 Weather Sensitive Hours (WSHs)

To facilitate generation planning and decision-making in the day-ahead and/or real-time energy market, accurate load forecast during peak times is particularly favored by the utilities for meeting peak demands and the implementation of peak shaving. As part of the load is used to maintain the ambient temperature and fulfill human comfort needs, the weekly/monthly/seasonal peak load is more likely to happen when the weather is either getting too warm or too cold.

FIGURE 14 shows a scatterplot of the load-temperature relationship for SA1 in the years 2013 and 2014. A black curve is fitted to observations. The red dashed lines highlighted the zone when the temperature is between 55°F and 70°F. The graph suggests that the load is generally at its lowest level when the temperature is within the 55~70°F region. In other words, the monthly/seasonal peak load is less likely to happen within this region. On the other hand, the load is generally higher outside of the 55~70°F region, suggesting a greater chance of reaching the peak. Besides, the forecasted load value (the black curve) is more sensitive to the deviations in the temperature input when the latter is outside of the 55~70°F region. This means that a smaller error in the temperature forecast may lead to a greater variation in the load forecast among the possible peak hours (which may create a more significant load forecast error as well). As our data is in hourly resolution, we denote the observations with a temperature forecast outside of the 55~70°F region as weather sensitive hours (WSHs). *FIGURE 15* shows the percentage ratio of WSHs within each month. The ratio of WSHs is higher during summer (July, August) and winter (December through February), and lower during the shoulder months.

To create a more meaningful and impactful analysis, we focus on predicting the temperature forecast quality among the WSHs and improving the load forecast accuracy for the future WSHs (i.e., hours having temperature forecast outside of the 55~70°F region). Concretely, we subset our data and keep only the WSHs when training and predicting the quality of the next day's temperature forecast.

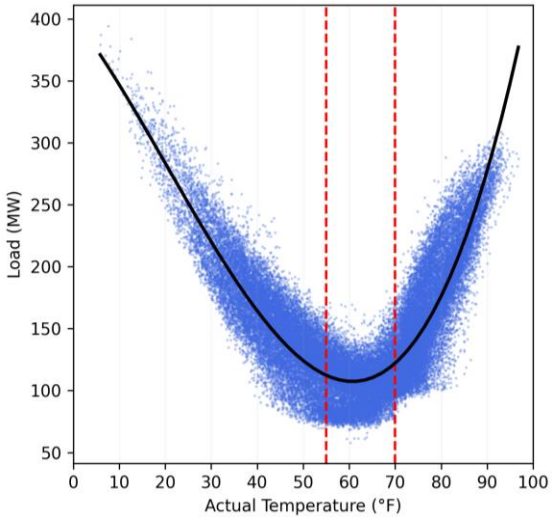

FIGURE 14: Load vs. temperature scatterplot with a fitted curve in black and red dashed lines showing temperature regions between 55°F and 70°F (SA1, years 2013 and 2014)

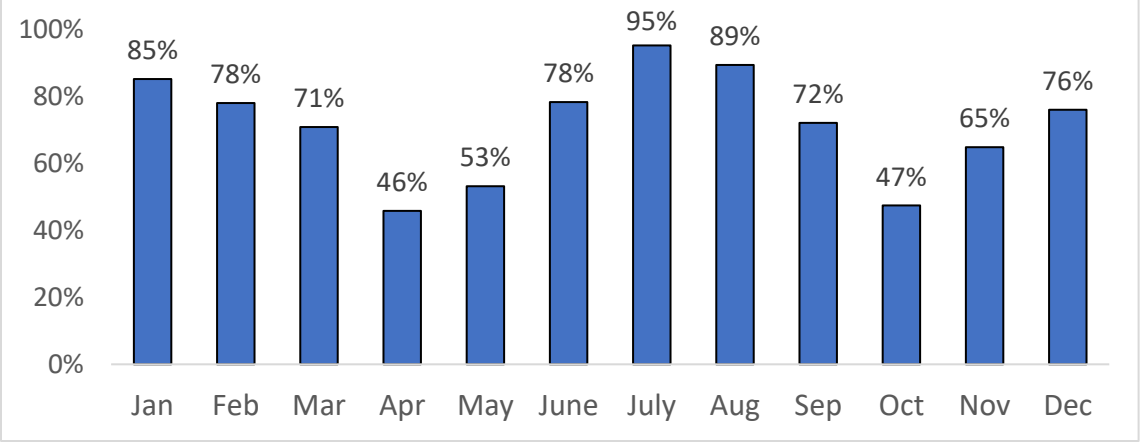

FIGURE 15: Percentage ratio of weather sensitive hours at each month (SA1, 2013 - 2018)

6.2. Methodology

6.2.1 Target Variable

In this study, our target variable is the absolute error of the day-ahead hourly composite temperature forecast. In other words, our forecast tells the quality (or accuracy) of a temperature forecast by providing its absolute error. We did not choose to forecast the *error* (which can be positive or negative) as we observe an inconsistent relationship between the *error* and the features we have chosen. The chosen features are detailed in Section 6.2.3. As our eventual goal is to leverage the prediction of this target variable for load forecasting, we assume that the under-forecast and over-forecast of the temperature have a similar impact on load forecast errors.

6.2.2 Error Metrics

We choose the mean absolute error (MAE) to measure the forecast accuracy of the target variable. This metric has been frequently used in the literature for temperature forecast evaluation (refer to Section 2.3). The MAE is defined as follows:

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

(14) – Error Measure: MAE

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

6.2.3 Features

An abrupt and steep variation in the temperature forecast (T_F) may suggest changes in the weather condition. Such changes could be caused by sunrise/sunset, seasonal changes, or some extreme weather events such as hurricanes and winter storms.

Depending on the speed and scale of the variations in the temperature forecast, the weather condition can be more challenging to predict in nature and thus results in a larger temperature forecast error.

FIGURE 16, *FIGURE 17*, and *FIGURE 18* show the scatter plots of the absolute error of temperature forecast (i.e., our target variable) with respect to the variations of T_F (within 1 to 9 hours), the daily average variation of T_F (within 1 to 9 hours), and diurnal (day-to-night) variation of T_F, respectively. The variations of T_F are all in absolute values. For instance, the variations of T_F within 2 hours at 9am is the absolute difference of the forecasted temperatures at 7am and 9am. The daily average variation of T_F within 2 hours is the daily (24-h) average of the variations of T_F within 2 hours. The diurnal (day-to-night) variation of T_F is the absolute difference between the maximum temperature and the minimum temperature of a calendar day (24 hours). A straight line is fitted to each subplot to denote the linear relationship between the two variables. The "m" values indicate the slope of the fitted lines. The denser color suggests that more observations are clustered within a region. In *FIGURE 16*, we observe that the target variable has a salient positive relationship with the variation of T_F within 1, 2, 3 and 4 hours. Besides, the positive relationship decays when the variation of T_F spans across longer hours (suggesting a slower change in T_F). Such a finding is due to the fact that the more abrupt variations in temperature are increasingly difficult to be precisely captured by the weather forecasting model. Similarly, in *FIGURE 17*, the target variable is positively related to the average temperature variation during that day, with a decaying positive relationship when the variation of T_F spans across longer hours. This implies that the days with more abrupt temperature variations are more likely to suffer from larger

forecasting errors. In *FIGURE 18*, we observe that the target variable is positively related to the diurnal variation of the temperature forecast. This implies that the day-to-night temperature variation can be another feature related to the target variable.

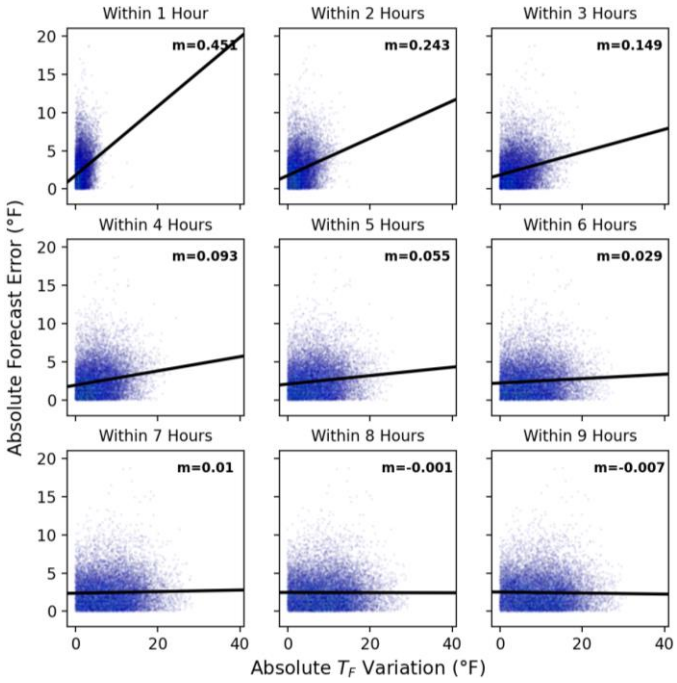

FIGURE 16: Scatterplots of day-ahead absolute temperature forecast error vs. temperature forecast variation (absolute) within 1 to 9 hours based on the training data (years 2013 and 2014)

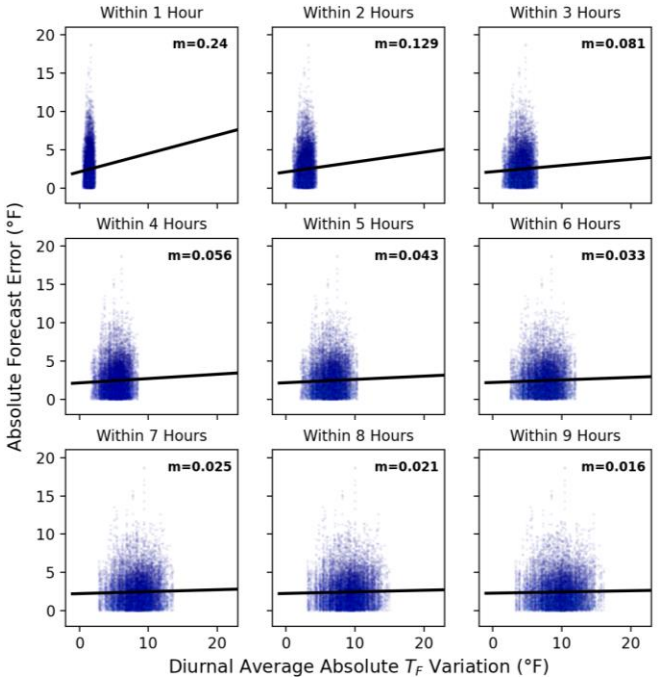

FIGURE 17: Scatterplots of day-ahead absolute temperature forecast error vs. daily (calendar) average temperature forecast variation (absolute) within 1 to 9 hours based on the training data (years 2013 and 2014)

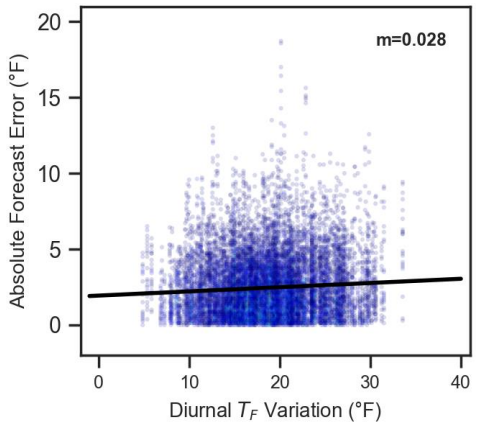

FIGURE 18: Scatterplot of day-ahead absolute temperature forecast error vs. diurnal (day-to-night for a calendar day) variation of temperature forecast (absolute) based on the training data (years 2013 and 2014)

FIGURE 19 shows a boxplot of the target variable at each hour of the day.

Noticeably, the errors are higher (at the median level, with a greater span) after sunrise (hours 7 to 11) and during sunset (hours 17 to 20). This is because the ambient

temperature gains a sharp increase after sunrise as the solar radiation heats the ground, and a sharp drop during the sunset. These fluctuations in temperature are natively challenging to capture. The plot also shows that the target variable has greater variations during the daytime (hours 7 to 20). During the night, the errors are generally lower, more consistent, and with fewer extreme errors (e.g., errors greater than 10°F). To forecast the target variable, we create categorical features for the hours during the day and group the hours into the same category during the nighttime.

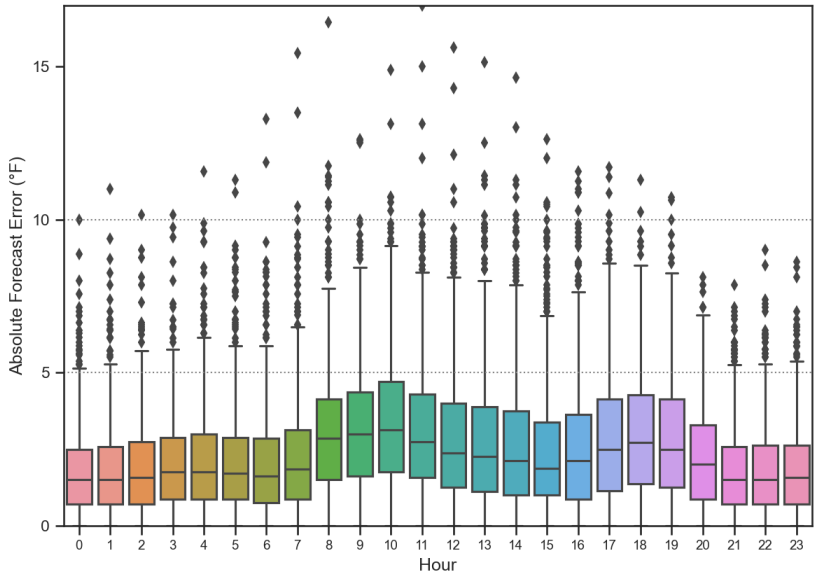

FIGURE 19: Boxplot of day-ahead absolute forecast error vs. hours-of-the-day based on the training data (years 2013 and 2014)

FIGURE 20 shows an ACF plot of the target variable time series. the target variable is strongly correlated to the first few lags. Besides, a correlation to the values 24 hours before the current hour can be observed. As we are forecasting the target variable up to 42 h into the future (our last available point in the historical data, i.e., the forecast origin, is 6am of the current day, and we are to forecast the target variable 24 h of the next day), we include a lagged dependent variable y_{t-48} in our model. We don't include

the y_{t-24} variable since the actual value of the target variable 24 h before some forecast hours is unknown.

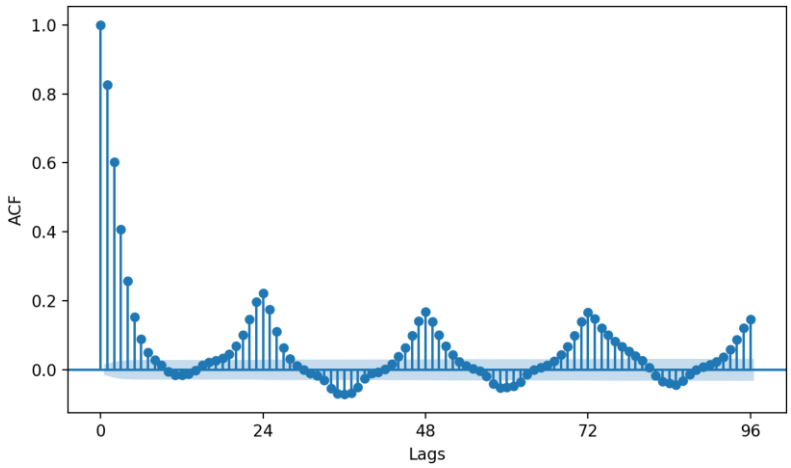

FIGURE 20: ACF plot of absolute temperature forecast errors on training data (years 2013 and 2014)

To sum up, the variables to be included in the modeling stage are listed as follows:

- Absolute variation of T_F within 1, 2, 3, and 4 hours
- Daily average absolute variation of T_F within 1, 2, 3, and 4 hours
- Diurnal (day-to-night) variation of T_F
- Hours-of-a-day categories (15 categories in total):
    a) Nighttime hours 0 - 6 and 21 - 23 are grouped into a single category
    b) The remaining hours are in their own categories (14 categories)
- A lagged dependent variable 48 hours before the forecast hours

6.2.4 Forecasting Models

This section provides descriptions of the forecasting models. We include three baseline models in this study:

- Persistence Naïve. The actual value at 6am of the current day is used to forecast the 24 hours of the next day.
- Seasonal Naïve. The forecast value equals its actual value 48 hours before that hour. In other words, the seasonal period equals 48 hours.
- Mean Naïve. The mean of historical values within the most recent 365 days is used to forecast the 24 hours of the next day.

These models are also known as the "naïve" models. By comparing their performances with the more advanced machine learning algorithms, we get to understand how the more complex model structure can benefit the forecasting performance.

The machine learning algorithms are listed as follows. For models ended with "by hour category", the data is sliced into 15 pieces based on the definition of the hours-of-a-day categories in Section 6.2.3, and the model training and validation are conducted separately within each hours-of-a-day subset. The remaining models introduce the hours-of-a-day categories as dummy variables.

- Multiple Linear Regression (MLR)
- LASSO
- ANN
- ANN (by hour category)
- Random Forest (RF)
- Random Forest (by hour category)
- XGBoost
- XGBoost (by hour category)

For models that require hyperparameter tuning (i.e., LASSO, ANN, Random Forest, and XGBoost), we refresh (re-tune) the hyperparameters once a year. Concretely, the hyperparameters used to forecast the validation year 2015 are tuned by using the year 2013 to fit the model and forecast onto the year 2014. Following a sliding simulation fashion as shown in *FIGURE 4*, The chosen set of hyperparameters (which minimizes the MAE in the year 2014) will be used to re-fit the model and generate day-ahead forecasts of the target variable in 2015. Thereafter, we repeat the same procedure to tune the hyperparameters used to forecast the rest years (2016, 2017, and 2018) and generate the day-ahead forecasts within each. *FIGURE 21* visualizes this procedure. For MLR and the baseline models that do not need hyperparameter tuning, we directly follow the sliding simulation procedure in *FIGURE 4* to generate the day-ahead forecasts for the years 2015 to 2018.

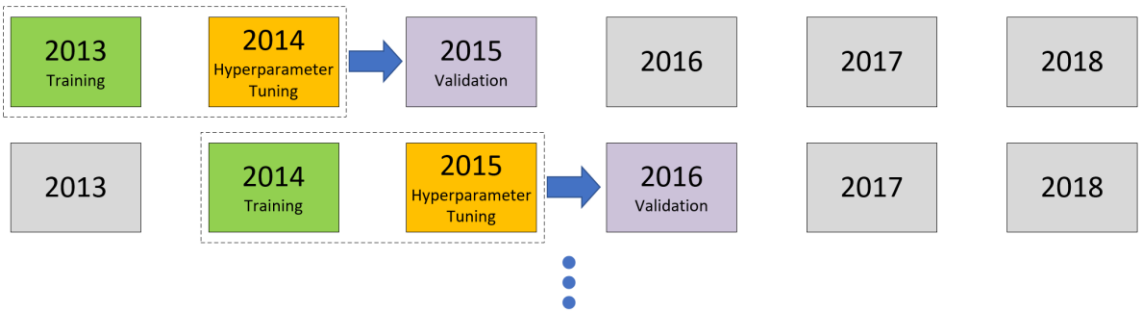

FIGURE 21: Hyperparameter tuning procedure for selected models

TABLE 10 summarizes the APIs being used and the search space of hyperparameters for the machine learning algorithms. For the hyperparameters that are not mentioned in the table, we go with the default values. Notably, we train the ANN models 10 times, each with a random weight initialization, and calculate the mean of the error metrics across the trials. More explanations on the hyperparameters can be found in

the official documentation of statsmodels 0.13 (Seabold & Perktold, 2010), TensorFlow 2.3.1 (Abadi et al., 2016), scikit-learn 1.0.2 (Pedregosa et al., 2011), and XGBoost 1.4 (T. Chen & Guestrin, 2016).

TABLE 10: API and hyperparameters of the machine learning algorithms

Method	API	Hyper-parameter	Search Space
MLR	statsmodels.api.OLS	-	-
LASSO	statsmodels.api.OLS	alpha	1e-5, 1e-4, 1e-3, 0.01, 0.1, 1
ANN	tensorflow.keras	activation	SELU
		batch_size	32
		No. of hidden layers	2
		No. of neurons in each layer	64
		Training Epochs	300 with Early Stopping
Random Forest	sklearn.ensemble. RandomForestRegressor	n_estimators	80 to 400 with a step of 20
		max_depth	8 to 30 with a step of 2
		min_samples_split	2 to 50 with a step of 5
		min_samples_leaf	1 to 50 with a step of 5
		max_features	0.2 to 0.9 with a step of 0.08
XGBoost	xgboost.XGBRegressor	n_estimators	50 to 300 with a step of 10
		max_depth	3 to 11 with a step of 2
		min_child_weight	1, 3, 5
		gamma	0, 1, 2, 3, 4
		subsample	0.6, 0.7, 0.8, 0.9
		colsample_bytree	0.6, 0.7, 0.8, 0.9
		reg_alpha	0, 0.001, 0.005, 0.01
		learning_rate	0.05, 0.1, 0.2

6.3. Results and Discussion

TABLE 11 shows a heatmap of the out-of-sample MAEs on the validation years for the prediction of temperature forecast errors under SA1. A greener color indicates a lower MAE value, while a redder color indicates a higher one. The bold values indicate the lowest errors of each year. Overall, no methods dominate in both validation years. Among the baseline models, the mean naïve method leads to the lowest MAEs in both validation years. Besides, all machine learning algorithms outperform the baseline models by a significant margin. The two ANN alternatives result in fairly similar performance. For SA1, the ANN produces good accuracy on both validation years and

results in the lowest MAE on average. Between the mean naïve method and ANN, the relative MAE reduction is 12% on average (i.e., from 1.407 to 1.234). Hence, we proceed with using ANN to generate the forecast for the test years (2017 and 2018). Similar findings were observed under SA2 (*TABLE 12*) and SA3 (*TABLE 13*), where the ANN (by hour category) and ANN were chosen for the two supply areas, respectively, to generate the forecasts for the test years.

TABLE 11: Heatmap of MAE values for temperature forecast error prediction based on the validation data (years 2015 and 2016), SA1

	Method	MAE (2015)	MAE (2016)	Overall MAE
Baseline models	Persistent Naïve	1.896	1.621	1.754
	Seasonal Naïve	1.940	1.565	1.746
	Mean Naïve	1.546	1.276	1.407
Machine Learning models	MLR	1.423	1.158	1.286
	LASSO	1.422	1.164	1.289
	ANN	**1.378**	1.099	**1.234**
	ANN (by hour category)	1.383	**1.097**	1.235
	RF	1.419	1.164	1.287
	RF by (by hour category)	1.434	1.167	1.296
	XGBoost	1.400	1.148	1.270
	XGBoost (by hour category)	1.423	1.150	1.282

TABLE 12: Heatmap of MAE values for temperature forecast error prediction based on the validation data (years 2015 and 2016), SA2

	Method	MAE (2015)	MAE (2016)	Overall MAE
Baseline models	Persistent Naïve	1.943	1.712	1.824
	Seasonal Naïve	1.955	1.616	1.780
	Mean Naïve	1.579	1.372	1.472
Machine Learning models	MLR	1.443	1.231	1.333
	LASSO	1.442	1.269	1.353
	ANN	1.382	**1.160**	1.2677
	ANN (by hour category)	**1.381**	1.161	**1.2676**
	RF	1.426	1.258	1.340
	RF by (by hour category)	1.433	1.231	1.329
	XGBoost	1.411	1.228	1.317
	XGBoost (by hour category)	1.425	1.224	1.321

TABLE 13: Heatmap of MAE values for temperature forecast error prediction based on the validation data (years 2015 and 2016), SA3

	Method	MAE (2015)	MAE (2016)	Overall MAE
Baseline models	Persistent Naïve	2.001	1.968	1.984
	Seasonal Naïve	2.013	1.868	1.937
	Mean Naïve	1.641	1.507	1.571
Machine Learning models	MLR	1.513	1.402	1.455
	LASSO	1.517	1.432	1.472
	ANN	**1.418**	1.340	**1.377**
	ANN (by hour category)	1.425	**1.336**	1.378
	RF	1.512	1.408	1.457
	RF by (by hour category)	1.510	1.391	1.448
	XGBoost	1.473	1.397	1.433
	XGBoost (by hour category)	1.500	1.376	1.435

FIGURE 22 shows the joint and marginal distribution of the temperature forecast error prediction against the actual during the test years (2017 and 2018). A black straight line is fitted to the scatter plot to denote the linear relationship between the prediction and the actual values. The "m" value indicates the slope of the fitted line. Overall, the prediction presents a salient positive relationship with the actual values, with a Pearson correlation coefficient of 0.477. This suggests that when the predicted temperature forecast error is higher, the actual temperature forecast error is generally higher. Taking a closer look at the density contours, most (absolute) temperature forecast errors are below 2 °F. Within this range, the prediction drifts slightly above the identity line, suggesting a tendency for over-forecasting. When the actual temperature error is above around 3 °F, the prediction generally drifts below the identity line, suggesting a tendency for under-forecasting.

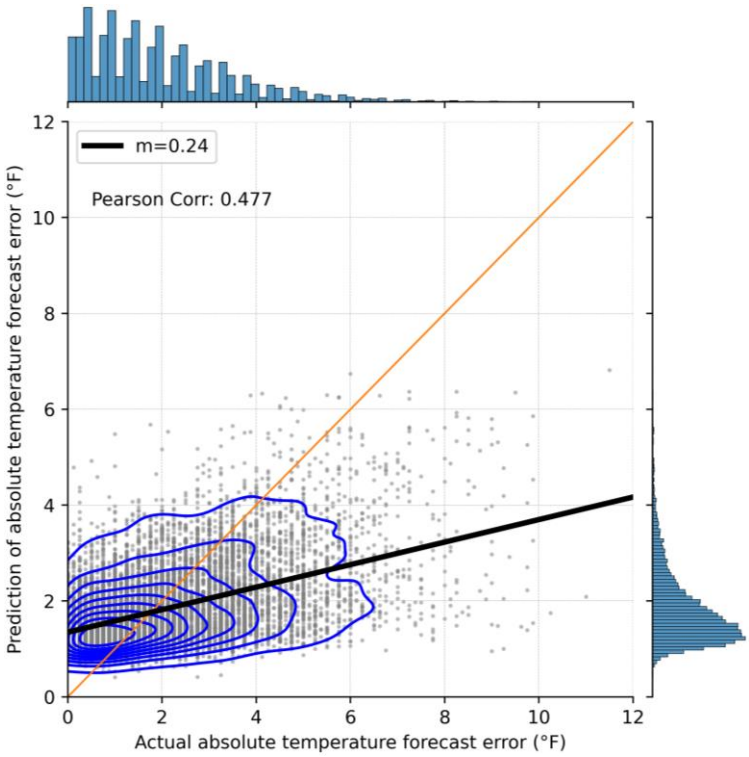

FIGURE 22: Joint and marginal distribution of temperature forecast error prediction against the actual under SA1 (2017-2018) with contour lines showing the kernel densities

FIGURE 23 shows some sample forecast output for SA1 using ANN among 20 random hours during the test years. When the actual temperature forecast error is lower (10 random hours, the upper plot), the model generally outputs lower predictions, with more prevalent over-forecasting. When the actual temperature forecast error is higher (10 random hours, the lower plot), the model generally outputs higher predictions, with prevalent under-forecasting. These findings align with the findings related to *FIGURE 22*, which provides further evidence that the forecasting model can capture the general trend of the temperature forecast quality. Similar findings were observed under the other two supply areas. To avoid verbose representation, we did not show them in this section.

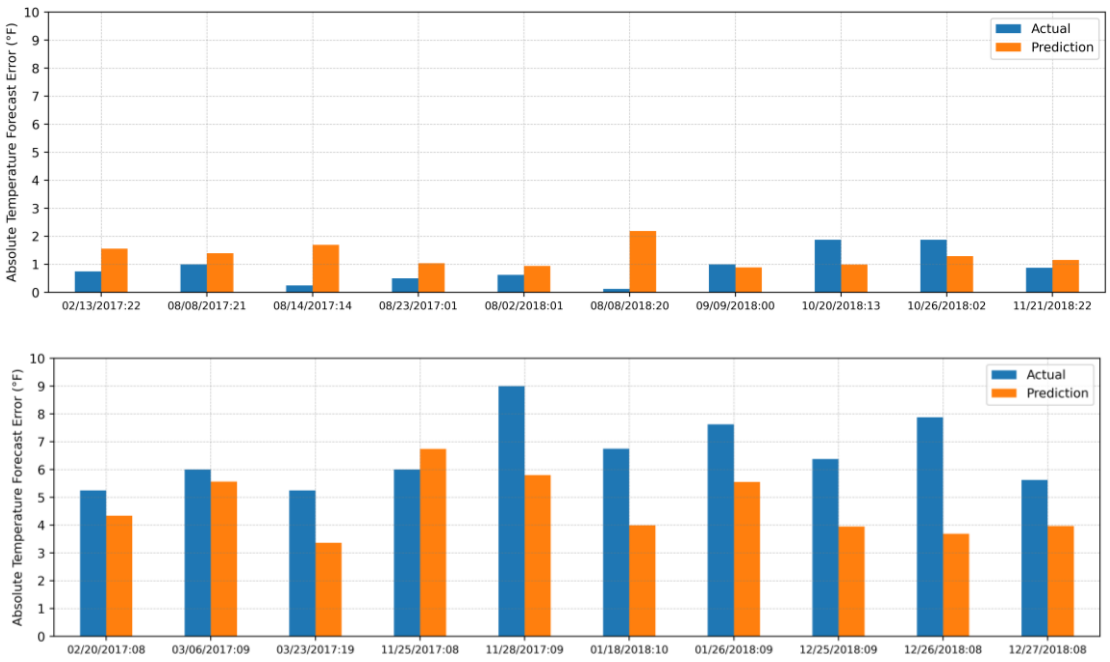

FIGURE 23: Bar charts of actual absolute temperature forecast error vs. predicted values for 20 random WSHs in the test years (2017 and 2018), when the actual values are lower (10 random WSHs, upper) and higher (10 random WSHs, lower), SA1

CHAPTER 7: MODEL SELECTION BASED ON TEMPERATURE FORECAST QUALITY PREDICTION

In this chapter, we propose a novel model selection methodology based on the predicted quality of temperature forecast as discussed in Chapter 6. We aim to improve the load forecast accuracy of the weather sensitive hours (WSHs) at diverse levels of temperature forecast accuracy. In the following sections, we start by introducing the overall procedure, our idea, and the forecasting models. Thereafter, we dive into the experiment results.

7.1. Methodology

7.1.1 Overall Procedure

FIGURE 24 shows a high-level workflow of the load forecasting procedure including a new model selection framework which will be detailed in this chapter. The raw data will first go through the steps of data cleansing and weather station selection. Thereafter, the data will be sliced into two portions. The WSHs will go through the module of temperature forecast quality prediction as introduced in Chapter 6. The output (i.e., "Temperature Forecast Error Prediction") will be used in the model selection step based on the ex-ante forecast accuracy of the validation data. We denote this new framework as M2. Two other frameworks discussed in Chapter 5 (i.e., M0 and M1) are used for comparison. The non-WSHs will follow the regular forecasting procedure by either using the ex-post (M0) or ex-ante (M1) forecast accuracy to select models and generate the forecast. The final product is a complete day-ahead load forecast. To form a

more impactful analysis, we focus on discussing the forecasting performance of the WSHs. The data partitioning rules in Chapter 5 are extended to this chapter.

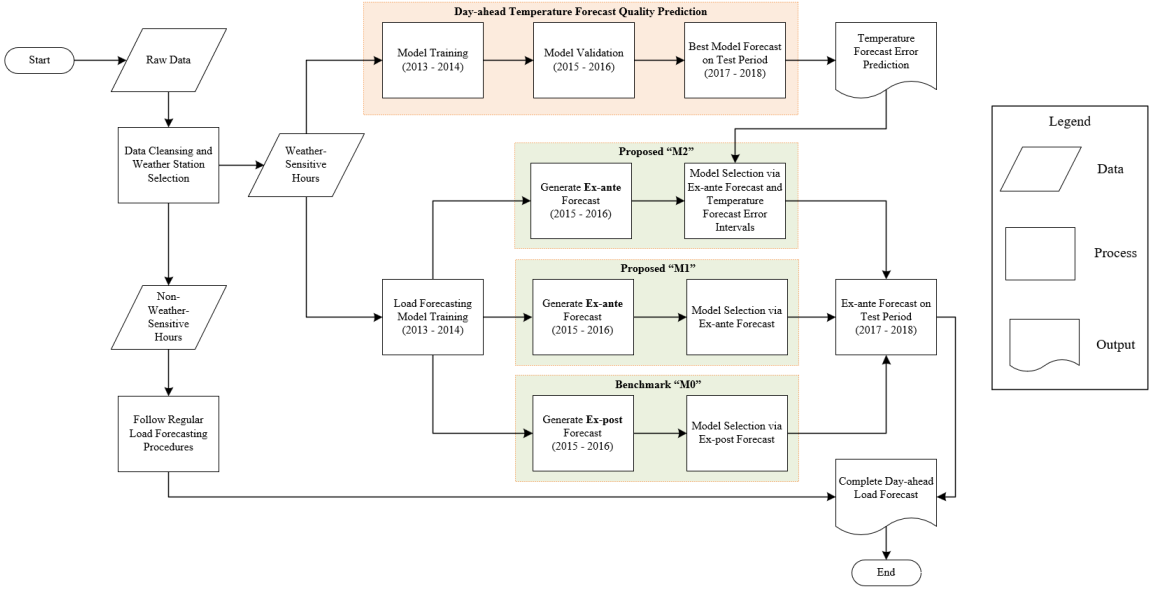

FIGURE 24: High-level workflow of the load forecasting procedure including model training, selection, and forecasting

7.1.2 General Idea

FIGURE 25 shows a histogram of the predicted temperature forecast error under SA1 in the years 2015 and 2016. The entire range of predicted values is divided into intervals with an equal size of 0.25 (for SA2 and SA3, we use an equal size of 0.35 due to a greater range of predicted values. To avoid verbose discussion, we did not visualize their distributions in this section). Our idea is straightforward – which is to select the load forecasting models within each predicted error interval based on the ex-ante load forecast accuracy. The intuition behind this is that the best-suited models can be different at each level (or interval) of the temperature forecast accuracy. Hence, a model selection process is required within each error interval. For the adjacent intervals that have too few WSHs

within, we group them into a single interval. For instance, in *FIGURE 25*, the error intervals less than 1°F are grouped. So are the ones above 5°F.

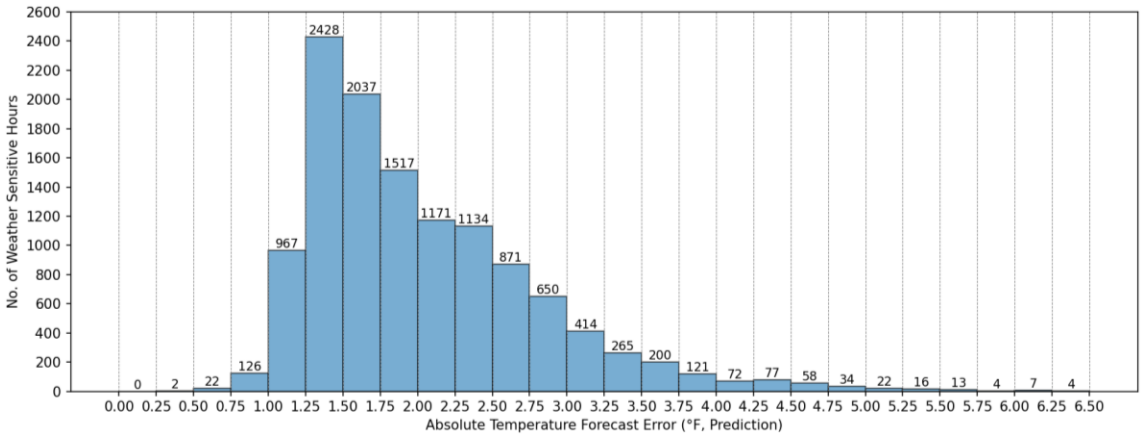

FIGURE 25: Histogram of the predicted temperature forecast error, SA1, validation years (2015 and 2016)

There are infinite ways to create intervals within a range of predicted temperature forecast errors. This can be an extra hyper-parameter to tune in practice. A rule of thumb is that an error interval should not be too wide (i.e., containing too many WSHs) or too narrow (i.e., containing too few WSHs). On one hand, if an interval is too wide, we may not fully leverage the power of this methodology. Think about an extreme case when placing all WSHs into the same interval – we did not get to leverage the temperature forecast quality information, and the proposed method would be equivalent to M1, which uses ex-ante load forecast accuracy for model selection. On the other hand, if an interval is too narrow, there is a potential overfitting issue. Think about an interval that contains only one or two WSHs on the validation data. There is a chance that tens or even hundreds of WSHs during the test period will be within the same interval. In this case, the model is selected by merely one or two WSHs, which may end up generalizing poorly on the test data. In this study, we ensure that each interval contains at least 15 WSHs. We

have not tried to break down the intervals with many WSHs within – we leave it as one of the future research directions.

7.1.3 Forecasting Models

In Chapter 5, a huge family of recency effect models has been presented based on the selection of (d, h) pairs. Due to the signal-error trade-off in the ex-ante forecasting settings (as discussed in Section 5.2.1), a better underlying model tends to have fewer lagged hourly temperatures (i.e., a smaller h) than the ones selected by the ex-post forecast accuracy. Hence, to avoid verbose presentation in this chapter, we only present a subset of recency effect models to demonstrate the proposed methodology. The candidate models are with $d = 1$ and have equal or fewer lagged hourly temperatures included than the ones selected by the ex-post forecast accuracy in Chapter 5.

For example, for SA1, the best recency effect model under the ex-post forecasting settings is $(d, h) = (1, 6)$ based on the validation data (refer to *TABLE 7*). In this chapter, we fix the value of $d = 1$ and discuss the models with decreasing h from 6 to 1. In practice, the search space of the (d, h) pairs can be adjusted based on the availability of the computational resources.

7.2. Results and Discussion

7.2.1 Model Selection Step

The day-ahead ex-ante load forecast is first generated on the validation years (2015 and 2016) using the candidate recency effect models. *TABLE 14* shows a heatmap of the results under SA1. A greener color indicates a lower ex-ante forecast MAPE value, while a redder color indicates a higher one. The “Intervals” column indicates the intervals of temperature forecast error prediction. The “No. of WSHs” column indicates the

number of WSHs within each error interval. The MAPE values (in %) are presented for each subset of WSHs. The bold values indicate the lowest MAPE values within each error interval. The corresponding recency effect model will be chosen to forecast the WSHs within the same error interval in the test period. At the bottom, the ex-ante and ex-post forecast MAPE values for all WSHs are provided. The ex-post forecast MAPE values are not color-coded since they are incomparable to the ex-ante forecast ones. Model (1, 1) is selected by the ex-ante forecast accuracy (the M1 framework), while model (1, 6) is selected by the ex-post forecast accuracy (the M0 framework).

TABLE 14: Heatmap of MAPE values (in %) based on temperature forecast error prediction, validation data (years 2015 and 2016), SA1

Intervals	No. of WSHs	Candidate Models (d, h)					
		(1, 6)	(1, 5)	(1, 4)	(1, 3)	(1, 2)	(1, 1)
[0, 1)	150	**4.648**	4.654	4.781	4.730	4.881	4.903
[1, 1.25)	967	5.739	5.750	5.707	5.729	**5.682**	5.720
[1.25, 1.5)	2428	6.155	6.098	6.061	6.010	5.954	**5.914**
[1.5, 1.75)	2037	6.427	6.350	6.260	6.208	6.179	**6.149**
[1.75, 2)	1517	7.014	6.957	6.883	6.803	6.796	**6.730**
[2, 2.25)	1171	7.432	7.339	7.288	7.125	7.108	**7.091**
[2.25, 2.5)	1134	7.516	7.464	7.372	7.307	**7.242**	7.244
[2.5, 2.75)	871	7.787	7.667	7.570	7.479	7.357	**7.235**
[2.75, 3)	650	7.451	7.389	7.327	7.381	7.321	**7.237**
[3, 3.25)	414	7.932	7.806	**7.727**	7.838	7.754	7.775
[3.25, 3.5)	265	7.171	7.138	7.073	**7.028**	7.157	7.356
[3.5, 3.75)	200	7.060	7.005	**6.908**	7.099	7.230	7.402
[3.75, 4)	121	6.871	6.790	**6.708**	6.951	7.128	7.096
[4, 4.25)	72	6.746	6.774	6.735	**6.447**	6.968	7.111
[4.25, 4.5)	77	6.167	6.110	**5.943**	5.993	6.429	6.981
[4.5, 4.75)	58	6.829	6.761	6.585	**6.361**	7.215	7.434
[4.75, 5)	34	7.747	7.580	7.368	**7.324**	7.863	8.459
[5, inf)	69	7.661	7.656	7.342	**6.746**	7.429	9.265
Ex-ante	12235	6.813	6.749	6.681	6.631	6.611	**6.604**
Ex-post	12235	**4.145**	4.158	4.176	4.202	4.241	4.314

TABLE 15 and *TABLE 16* show the results on the validation years under SA2 and SA3, respectively, following the same format as *TABLE 14*. At the bottom of each table, it shows that the model (1, 3) and (1, 10) are selected by the ex-ante (M1) and ex-post

(M0) forecast accuracy under SA2, respectively. Model (1, 1) and (1, 10) are selected by the ex-ante and ex-post forecast accuracy under SA3, respectively.

TABLE 15: Heatmap of MAPE values (in %) based on temperature forecast error prediction, validation data (years 2015 and 2016), SA2

Intervals	No. of WSHs	Candidate Models (d, h)									
		(1, 10)	(1, 9)	(1, 8)	(1, 7)	(1, 6)	(1, 5)	(1, 4)	(1, 3)	(1, 2)	(1, 1)
[0, 1)	282	6.700	6.593	6.481	6.304	6.105	5.948	5.861	5.778	5.735	**5.634**
[1, 1.35)	1849	5.778	5.806	5.788	5.715	5.676	5.653	5.601	5.547	5.512	**5.476**
[1.35, 1.7)	3086	6.070	6.056	6.026	5.994	5.955	5.907	5.864	5.828	5.778	**5.740**
[1.7, 2.05)	1882	6.739	6.781	6.686	6.576	6.519	6.484	6.433	6.393	6.408	**6.380**
[2.05, 2.4)	1491	6.784	6.805	6.724	6.664	6.582	6.566	6.536	6.509	6.496	**6.442**
[2.4, 2.75)	1259	7.357	7.314	7.236	7.228	7.203	7.170	7.186	7.104	7.065	**6.925**
[2.75, 3.1)	886	7.496	7.404	7.422	7.410	7.492	7.385	**7.317**	7.365	7.464	7.459
[3.1, 3.45)	610	7.579	7.463	7.531	7.573	7.513	7.402	**7.273**	7.340	7.392	7.559
[3.45, 3.8)	370	7.702	**7.604**	7.714	7.784	7.765	7.741	7.611	7.737	7.929	8.164
[3.8, 4.15)	201	7.970	7.737	7.708	7.720	7.574	7.454	**7.217**	7.287	7.435	7.526
[4.15, 4.5)	118	8.575	8.166	8.224	8.337	8.243	8.058	**7.807**	8.121	8.452	8.997
[4.5, 4.85)	71	9.124	8.205	8.105	8.102	7.911	**7.581**	7.586	7.628	7.967	8.344
[4.85, 5.2)	48	10.017	9.515	9.400	9.377	9.154	8.785	8.360	**8.348**	9.154	9.220
[5.2, 5.55)	30	9.804	9.188	9.662	9.637	9.220	8.801	8.372	**8.172**	8.457	9.650
[5.55, 5.9)	22	6.736	6.277	6.481	6.470	6.442	6.268	6.280	**5.765**	6.061	7.140
[5.9, inf)	46	8.718	8.696	8.934	8.895	8.665	8.499	8.104	**7.636**	8.358	11.436
Ex-ante	12251	6.699	6.669	6.633	6.589	6.543	6.489	6.433	**6.408**	6.416	6.414
Ex-post	12251	**3.324**	3.329	3.332	3.335	3.332	3.340	3.353	3.385	3.444	3.553

TABLE 16: Heatmap of MAPE values (in %) based on temperature forecast error prediction, validation data (years 2015 and 2016), SA3

Intervals	No. of WSHs	Candidate Models (d, h)										
		(1, 11)	(1, 10)	(1, 9)	(1, 8)	(1, 7)	(1, 6)	(1, 5)	(1, 4)	(1, 3)	(1, 2)	(1, 1)
[0, 1)	41	6.238	6.120	6.235	6.359	6.418	6.173	6.105	**5.968**	6.092	6.451	6.503
[1, 1.35)	485	5.875	5.858	5.797	5.742	5.729	5.748	5.689	5.641	5.649	**5.563**	5.572
[1.35, 1.7)	2341	5.820	5.830	5.850	5.842	5.834	5.807	5.786	5.754	5.689	5.688	**5.649**
[1.7, 2.05)	2948	6.320	6.317	6.311	6.303	6.283	6.289	6.263	6.212	6.142	6.094	**6.027**
[2.05, 2.4)	2059	6.561	6.566	6.547	6.549	6.517	6.522	6.578	6.550	6.475	6.476	**6.401**
[2.4, 2.75)	1456	7.434	7.454	7.424	7.381	7.354	7.361	7.376	7.419	7.372	7.310	**7.177**
[2.75, 3.1)	1104	7.040	7.017	6.930	6.865	6.839	6.797	**6.782**	6.931	7.053	6.960	6.935
[3.1, 3.45)	725	7.649	7.608	7.527	7.466	7.348	7.327	**7.277**	7.283	7.653	7.702	7.705
[3.45, 3.8)	405	7.428	7.348	7.259	7.228	7.167	7.058	**6.931**	6.969	7.288	7.513	7.691
[3.8, 4.15)	262	8.164	8.016	7.798	7.851	7.891	7.685	7.668	**7.587**	7.817	8.018	8.327
[4.15, 4.5)	144	8.120	7.714	7.461	7.644	7.680	7.443	**7.342**	7.409	7.748	8.001	8.312
[4.5, 4.85)	84	7.845	7.734	7.523	7.671	7.620	7.396	**7.383**	7.493	7.505	7.865	8.194
[4.85, 5.2)	42	6.492	6.192	6.239	6.399	6.385	6.197	6.082	**5.914**	5.977	6.049	7.218
[5.2, 5.55)	21	6.046	5.790	5.974	5.869	5.693	5.740	5.574	5.784	**5.436**	5.528	7.282
[5.55, 5.9)	15	6.890	6.951	7.108	7.091	7.052	6.976	6.957	6.829	**6.557**	7.361	8.847
[5.9, inf)	28	7.404	**6.959**	7.182	7.335	7.411	7.302	7.401	7.509	7.354	7.754	8.763
Ex-ante	12160	6.636	6.621	6.591	6.576	6.549	6.529	6.517	6.511	6.516	6.508	**6.484**
Ex-post	12160	3.780	**3.777**	3.779	3.786	3.792	3.798	3.813	3.838	3.878	3.932	4.023

7.2.2 Out-of-sample Test

The models selected based on the validation data are used to generate the forecast for the test years (2017 and 2018). *TABLE 17* shows a heatmap of the test years' performance under SA1. The "M0: (1, 6)", "M1: (1, 1)", and "M2" columns provide the MAPE values (in %) based on the model(s) selected by each method. A green and bold cell indicates the lowest MAPE value of a specific error interval among the three methods. A yellow cell highlights the 2nd lowest MAPE value, while a red cell highlights the highest one. When two methods select the same model within an error interval, their MAPE values will be identical and end up having the same color.

TABLE 17: Heatmap of MAPE values (in %) based on temperature forecast error prediction, test data (years 2017 and 2018), SA1

Intervals	No. of WSHs	M0: (1, 6)	M1: (1, 1)	M2
[0, 1)	635	5.574	**5.487**	5.574
[1, 1.25)	2280	4.993	4.852	**4.848**
[1.25, 1.5)	2575	5.461	**5.344**	**5.344**
[1.5, 1.75)	2089	6.011	**5.689**	**5.689**
[1.75, 2)	1550	6.754	**6.389**	**6.389**
[2, 2.25)	1061	7.078	**6.788**	**6.788**
[2.25, 2.5)	705	6.951	6.725	**6.639**
[2.5, 2.75)	496	6.968	**6.797**	**6.797**
[2.75, 3)	386	**7.105**	7.322	7.322
[3, 3.25)	321	7.994	7.873	**7.768**
[3.25, 3.5)	224	8.462	**8.053**	8.271
[3.5, 3.75)	156	7.231	**6.900**	7.120
[3.75, 4)	107	8.023	7.782	**7.563**
[4, 4.25)	78	8.415	8.227	**7.819**
[4.25, 4.5)	39	8.608	8.571	**8.268**
[4.5, 4.75)	40	9.223	9.290	**8.047**
[4.75, 5)	36	9.541	9.553	**8.146**
[5, inf)	76	9.358	10.275	**8.459**
All WSHs	12854	6.183	5.986	**5.965**
WSHs in [0, 3.75)	12478	6.107	**5.902**	5.905
WSHs in [3.75, inf)	376	8.708	8.790	**7.978**

Between the M0 and M1 alternatives, M1 mostly leads to superior accuracy when the temperature forecast error prediction is below 4.5 °F. Nevertheless, it results in higher

MAPEs when the temperature forecast error prediction is above 4.5 °F. On the other hand, the M2 framework leads to the best forecast accuracy among the three alternatives for the vast majority of the error intervals. This is because the M2 framework grants the flexibility to choose different models at diverse levels of temperature forecast accuracy.

The MAPE values for all WSHs, the WSHs within a lower error range (i.e., error intervals $<$ 3.75 °F), and a higher error range (i.e., error intervals $\geq$ 3.75 °F) are provided at the bottom of *TABLE 17*, respectively. For all WSHs, the M1 framework outperforms the M0 counterpart with a relative MAPE reduction of 3.2% (from 6.183% to 5.986%). The M2 framework improves further on top of the M1 counterpart. For the WSHs within the error interval of [0, 3.75), M2 leads to similar performance as M1, while both outperform M0 significantly. For the 376 WSHs within the error interval of [3.75, inf), the M2 framework leads to significant MAPE improvement: compared to the M0 and M1 counterparts, the relative MAPE reduction by using the M2 framework is 8.4% (from 8.708% to 7.978%) and 9.2% (from 8.790% to 7.978%), respectively.

TABLE 18 and *TABLE 19* provide the test years' performance for SA2 and SA3, respectively. Similar findings can be observed other than a few exceptions. For SA2, the M1 framework already provides promising improvement over the M0 counterpart within almost all error intervals, while the M2 framework extends the improvement and leads to even better overall performance. For SA3, between the M0 and M1 alternatives, the M1 framework leads to promising accuracy when the predicted temperature forecast error is below 2.4 °F. However, the overall improvement among all WSHs is marginal due to M1's underperformance at the higher error intervals. On the other hand, the M2 framework leads to superior accuracy than the M1 counterpart at the vast majority of the

error intervals. Compared to the M0 and M1 counterparts, the relative MAPE reduction overall by using the M2 framework is 1.9% (from 6.222% to 6.104%) and 1.7% (from 6.212% to 6.104%), respectively.

To summarize the overall performance of all three supply areas, the M1 framework consistently outperforms the M0 counterpart among all WSHs, with an average improvement in accuracy of 3.1%, whereas the M1 framework may lead to poor performance at higher error intervals. On the other hand, the M2 framework introduces extensive flexibility in choosing suitable models at diverse levels of temperature forecast accuracy. At lower error intervals, the performance of M2 is similar to or better than M1. At higher error intervals, the M2 framework makes up for the underperformance of M1 by a 7.4% improvement on average. Among the three model selection frameworks, our case study shows that the M2 framework leads to the best accuracy overall across all three supply areas. Among all weather sensitive hours, the M2 framework leads to an average of 0.8% improvement over the M1 framework and 3.9% improvement over the M0 benchmark.

TABLE 18: Heatmap of MAPE values (in %) based on temperature forecast error prediction, test data (years 2017 and 2018), SA2

Intervals	No. of WSHs	M0: (1, 10)	M1: (1, 3)	M2
[0, 1)	850	6.106	5.484	**5.432**
[1, 1.35)	3391	5.309	4.981	**4.880**
[1.35, 1.7)	2734	6.178	5.712	**5.705**
[1.7, 2.05)	1826	6.354	**5.877**	5.938
[2.05, 2.4)	1269	6.995	**6.576**	6.715
[2.4, 2.75)	759	6.949	**6.542**	6.810
[2.75, 3.1)	559	7.209	6.985	**6.862**
[3.1, 3.45)	483	7.865	7.823	**7.691**
[3.45, 3.8)	290	7.874	7.803	**7.503**
[3.8, 4.15)	163	**7.436**	7.452	7.461
[4.15, 4.5)	90	8.357	8.669	**8.355**
[4.5, 4.85)	52	7.502	**7.285**	7.653
[4.85, 5.2)	37	8.000	**7.674**	**7.674**
[5.2, 5.55)	35	**8.576**	8.598	8.598
[5.55, 5.9)	17	11.566	**11.383**	**11.383**
[5.9, inf)	35	11.370	**10.867**	**10.867**
All WSHs	12590	6.314	5.932	**5.921**
WSHs in [0, 3.8)	12161	6.245	5.851	**5.840**
WSHs in [3.8, inf)	429	8.264	8.234	**8.216**

TABLE 19: Heatmap of MAPE values (in %) based on temperature forecast error prediction, test data (years 2017 and 2018), SA3

Intervals	No. of WSHs	M0: (1, 10)	M1: (1, 3)	M2
[0, 1)	131	6.881	**6.785**	6.841
[1, 1.35)	1058	5.645	**5.159**	5.268
[1.35, 1.7)	3372	5.644	**5.380**	**5.380**
[1.7, 2.05)	2813	5.825	**5.663**	**5.663**
[2.05, 2.4)	1565	6.383	**6.307**	**6.307**
[2.4, 2.75)	1021	**6.936**	6.988	6.988
[2.75, 3.1)	753	**7.302**	7.588	7.441
[3.1, 3.45)	605	**7.172**	7.752	7.341
[3.45, 3.8)	474	**7.203**	7.855	7.312
[3.8, 4.15)	317	**7.199**	8.015	7.339
[4.15, 4.5)	183	7.098	7.851	**7.090**
[4.5, 4.85)	124	6.562	7.480	**6.514**
[4.85, 5.2)	68	**7.344**	9.482	7.488
[5.2, 5.55)	61	**7.612**	9.375	8.034
[5.55, 5.9)	52	6.722	8.926	**6.714**
[5.9, inf)	26	**8.165**	10.791	**8.165**
All WSHs	12623	6.222	6.212	**6.104**
WSHs in [0, 3.8)	11792	6.158	6.068	**6.026**
WSHs in [3.8, inf)	831	**7.124**	8.263	7.211

CHAPTER 8: CONCLUSION

Electric load forecasting is an integral part of the modern power system. The field of STLF has been extensively studied by industry and research groups as power utilities need accurate short-term load forecasts for better decision-making in their daily operations. Most existing literature build and test load forecasting models under the ex-post forecasting settings, where the actual weather information is used in the forecast period. Nevertheless, the robustness of these models under the operational (ex-ante) forecasting settings has rarely been studied, where the impact of imperfect weather information on the model has to be considered. This status quo is often due to the shortage of historical weather forecasts in the model development stage.

This dissertation aims to close this gap by presenting two new model selection frameworks for selecting better models in STLF. The first framework (M1) selects models based on the historical ex-ante load forecast accuracy. Existing literature suggests that the ex-ante load forecast accuracy can be impacted by the levels of temperature forecast accuracy. Therefore, as an extension to the M1 framework, the second framework (M2) is proposed to select models based on the historical ex-ante load forecast accuracy at diverse levels of temperature forecast accuracy. Since the temperature forecast accuracy is unknown ahead of time, we propose a novel solution to predict the day-ahead temperature forecast errors and compared the prediction accuracy among multiple baselines and machine learning methods. The prediction output shows promising performance in capturing the overall trend of temperature forecast accuracy. The temperature forecast error prediction then serves as an input to the model selection step of the M2 framework to enable model selection within each level of the temperature forecast error prediction. The two

proposed frameworks have been compared to a benchmark framework (M0), which follows the conventional practice in the literature by using historical ex-post load forecast accuracy for model selection.

The proposed solutions offer practical values in field operations. To create a more impactful analysis, the performances of the three model selection frameworks are compared among the weather sensitive hours, when the load is more likely to reach the monthly or seasonal peak, and a smaller error in the temperature forecast may lead to a greater inaccuracy in the load forecast. Through an empirical case study at a medium-sized US utility, results show that the day-ahead load forecast accuracy can be significantly improved by the two proposed frameworks in comparison to the benchmark framework. Besides, the M2 framework consistently achieves the best overall accuracy across all three supply areas. The superiority of the M2 framework suggests that diversified models built for different scenarios may be favored to achieve better load forecast accuracy. Computational-wise, the M1 framework leads to the same computational cost as the M0 framework. The M2 framework unlocks further improvement in forecast accuracy with additional computational cost in temperature forecast error prediction.

This research also sparks a few future research directions. First, the candidate models in this research are based on the recency effect modeling framework investigated in (P. Wang et al., 2016). As discussed in Section 3.3, this framework follows an incremental manner of introducing the preceding hourly temperatures, which lacks the flexibility to drop certain temperatures when the predicted quality of these temperatures gets worse (i.e., with higher errors). Therefore, additional model structures can be tested to selectively pick the recent temperatures that are believed to have better quality. Second, as

discussed in Section 7.1.2, there are infinite ways to define the error intervals based on the temperature forecast error prediction and select suitable models within each. Future research may explore additional ways to leverage temperature forecast error prediction for model selection. Third, the weather variable in this research is limited to temperatures. As other weather variables such as relative humidity and wind speed may further improve the load forecast accuracy (refer to (Xie et al., 2018) and (Xie & Hong, 2017), respectively), additional research may investigate the efficacy of the proposed frameworks with additional weather variables involved. The quality of these weather variables may be predicted following the thought process as presented in Section 6.2. Fourth, the case study in this research focuses on day-ahead forecasting. Future research may extend the proposed frameworks to additional lead times.

REFERENCES

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., … Zheng, X. (2016). TensorFlow: A system for large-scale machine learning. *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016*, 265–283. https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi

Agüera-Pérez, A., Palomares-Salas, J. C., González de la Rosa, J. J., & Florencias-Oliveros, O. (2018). Weather forecasts for microgrid energy management: Review, discussion and recommendations. *Applied Energy*, *228*, 265–278. https://doi.org/10.1016/j.apenergy.2018.06.087

Alasali, F., Nusair, K., Alhmoud, L., & Zarour, E. (2021). Impact of the COVID-19 Pandemic on Electricity Demand and Load Forecasting. *Sustainability*, *13*(3), 1435. https://doi.org/10.3390/su13031435

Amjady, N. (2001). Short-term hourly load forecasting using time-series modeling with peak load estimation capability. *IEEE Transactions on Power Systems*, *16*(3), 498–505. https://doi.org/10.1109/59.932287

Chapagain, K., & Kittipiyakul, S. (2018). Performance Analysis of Short-Term Electricity Demand with Atmospheric Variables. *Energies*, *11*(4), 818. https://doi.org/10.3390/en11040818

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

Charney, J. G., FjÖrtoft, R., & Neumann, J. Von. (1950). Numerical Integration of the Barotropic Vorticity Equation. *Tellus*, *2*(4), 237–254. https://doi.org/10.3402/tellusa.v2i4.8607

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

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

Chen, S. T., Yu, D. C., & Moghaddamjo, A. R. (1992). Weather Sensitive Short-Term Load Forecasting Using Nonfully Connected Artificial Neural Network. *IEEE*

Transactions on Power Systems, *7*(3), 1098–1105. https://doi.org/10.1109/59.207323

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2939672.2939785

Cheng, W. Y. Y., & Steenburgh, W. J. (2005). Evaluation of surface sensible weather forecasts by the WRF and the Eta Models over the western United States. *Weather and Forecasting*, *20*(5), 812–821. https://doi.org/10.1175/WAF885.1

Chitalia, G., Pipattanasomporn, M., Garg, V., & Rahman, S. (2020). Robust short-term electrical load forecasting framework for commercial buildings using deep recurrent neural networks. *Applied Energy*, *278*, 115410. https://doi.org/10.1016/J.APENERGY.2020.115410

Dahl, M., Brun, A., Kirsebom, O. S., & Andresen, G. B. (2018). Improving Short-Term Heat Load Forecasts with Calendar and Holiday Data. *Energies 2018, Vol. 11, Page 1678*, *11*(7), 1678. https://doi.org/10.3390/EN11071678

De Felice, M., Alessandri, A., & Ruti, P. M. (2013). Electricity demand forecasting over Italy: Potential benefits using numerical weather prediction models. *Electric Power Systems Research*, *104*, 71–79. https://doi.org/10.1016/J.EPSR.2013.06.004

Douglas, A. P., Breipohl, A. M., Lee, F. N., & Adapa, R. (1998). The impacts of temperature forecast uncertainty on bayesian load forecasting. *IEEE Transactions on Power Systems*, *13*(4), 1507–1513. https://doi.org/10.1109/59.736298

Fan, C., Liao, Y., Zhou, G., Zhou, X., & Ding, Y. (2020). Improving cooling load prediction reliability for HVAC system using Monte-Carlo simulation to deal with uncertainties in input variables. *Energy and Buildings*, *226*, 110372. https://doi.org/10.1016/J.ENBUILD.2020.110372

Fan, S., Chen, L., & Lee, W. J. (2008). Short-term load forecasting using comprehensive combination based on multi- meteorological information. *Conference Record - Industrial and Commercial Power Systems Technical Conference*. https://doi.org/10.1109/ICPS.2008.4606288

Fan, Shu, & 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

Fan, Shu, 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

Fay, D., & Ringwood, J. V. (2010). On the influence of weather forecast errors in short-term load forecasting models. *IEEE Transactions on Power Systems*, *25*(3), 1751–

1758. https://doi.org/10.1109/TPWRS.2009.2038704

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

Goude, Y., Nedellec, R., & Kong, N. (2014). Local short and middle term electricity load forecasting with semi-parametric additive models. *IEEE Transactions on Smart Grid*, *5*(1), 440–446. https://doi.org/10.1109/TSG.2013.2278425

Grönås, S. (1985). *A pilot study on the prediction of medium range forecast quality*. *119*, 22. https://doi.org/10.21957/ostzejo17

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

Haben, S., Giasemidis, G., Ziel, F., & Arora, S. (2019). Short term load forecasting and the effect of temperature at the low voltage level. *International Journal of Forecasting*, *35*(4), 1469–1484. https://doi.org/10.1016/j.ijforecast.2018.10.007

Hagan, M. T., & Behr, S. M. (1987). The Time Series Approach to Short Term Load Forecasting. *IEEE Transactions on Power Systems*, *2*(3), 785–791. https://doi.org/10.1109/TPWRS.1987.4335210

Hansen, J. W. (2002). Realizing the potential benefits of climate prediction to agriculture: Issues, approaches, challenges. *Agricultural Systems*. https://doi.org/10.1016/S0308-521X(02)00043-4

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

Hoffman, R. N., Boukabara, S. A., Kumar, V. K., Garrett, K., Casey, S. P. F., & Atlas, R. (2017). An Empirical cumulative density function approach to defining summary NWP forecast assessment metrics. *Monthly Weather Review*, *145*(4), 1427–1435. https://doi.org/10.1175/MWR-D-16-0271.1

Hoffman, R. N., Kumar, V. K., Boukabara, S. A., Ide, K., Yang, F., & Atlas, R. (2018). Progress in forecast skill at three leading global operational NWP centers during 2015-17 as seen in summary assessment metrics (SAMs). *Weather and Forecasting*, *33*(6), 1661–1679. https://doi.org/10.1175/WAF-D-18-0117.1

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

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

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

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

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

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

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

Hong, T., Wilson, J., & Xie, J. (2014). Long term probabilistic load forecasting and normalization with hourly information. *IEEE Transactions on Smart Grid*, *5*(1), 456–462. https://doi.org/10.1109/TSG.2013.2274373

Hong, T., Xie, J., & Black, J. (2019). Global energy forecasting competition 2017: Hierarchical probabilistic load forecasting. *International Journal of Forecasting*, *35*(4), 1389–1399. https://doi.org/10.1016/j.ijforecast.2019.02.006

Hyndman, R., Koehler, A., Ord, K., & Snyder, R. (2008). *Forecasting with Exponential Smoothing: The State Space Approach*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-71918-2

Jolliffe, I. T., & Stephenson, D. B. (2012). *Forecast Verification: A Practioner's Guide in Atmospheric Science* (2nd ed.). John Wiley & Sons. https://doi.org/10.1002/9781119960003

Kalnay, E. (2019). Historical perspective: earlier ensembles and forecasting forecast skill. *Quarterly Journal of the Royal Meteorological Society*, *145*(S1), 25–34. https://doi.org/10.1002/qj.3595

Kalnay, E., & Dalcher, A. (1987). Forecasting forecast skill. *Monthly Weather Review*, *115*(2), 349–356. https://doi.org/10.1175/1520-0493(1987)115<0349:FFS>2.0.CO;2

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

Kulkarni, S., Simon, S. P., & Sundareswaran, K. (2013). A spiking neural network (SNN) forecast engine for short-term electrical load forecasting. *Applied Soft Computing Journal*, *13*(8), 3628–3635. https://doi.org/10.1016/j.asoc.2013.04.007

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.net/public/sites/energy/reports/electricloadforecasting.pdf

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

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

Luo, J., Hong, T., & Fang, S. C. (2018a). Benchmarking robustness of load forecasting models under data integrity attacks. *International Journal of Forecasting*, *34*(1). https://doi.org/10.1016/j.ijforecast.2017.08.004

Luo, J., Hong, T., & Fang, S. C. (2018b). Robust Regression Models for Load Forecasting. *IEEE Transactions on Smart Grid*, *10*(5), 5397–5404. https://doi.org/10.1109/TSG.2018.2881562

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

Lusis, P., Khalilpour, K. R., Andrew, L., & Liebman, A. (2017). Short-term residential load forecasting: Impact of calendar effects and forecast granularity. *Applied Energy*, *205*, 654–669. https://doi.org/10.1016/j.apenergy.2017.07.114

Mandal, P., Senjyu, T., Urasaki, N., & Funabashi, T. (2006). A neural network based several-hour-ahead electric load forecasting using similar days approach. *International Journal of Electrical Power & Energy Systems*, *28*(6), 367–373. https://doi.org/10.1016/J.IJEPES.2005.12.007

Methaprayoon, K., Lee, W. J., Rasmiddatta, S., Liao, J. R., & Ross, R. J. (2007). Multistage artificial neural network short-term load forecasting engine with front-end weather forecast. *IEEE Transactions on Industry Applications*, *43*(6), 1410–1416. https://doi.org/10.1109/TIA.2007.908190

Molteni, F., & Palmer, T. N. (1991). A real-time scheme for the prediction of forecast

skill. *Monthly Weather Review*, *119*(4), 1088–1097. https://doi.org/10.1175/1520-0493(1991)119<1088:ARTSFT>2.0.CO;2

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*, *36*(2), 695–712. https://doi.org/10.1016/j.ijforecast.2019.08.008

Murphy, A. H. (1988). Skill scores based on the mean square error and their relationships to the correlation coefficient. *Monthly Weather Review*, *116*(12), 2417–2424. https://doi.org/10.1175/1520-0493(1988)116<2417:SSBOTM>2.0.CO;2

Murphy, A. H. (1993). What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather & Forecasting*, *8*(2), 281–293. https://doi.org/10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2

Murphy, A. H., & Winkler, R. L. (1987). A general framework for forecast verification. *Monthly Weather Review*, *115*(7), 1330–1338. https://doi.org/10.1175/1520-0493(1987)115<1330:AGFFFV>2.0.CO;2

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

Neto, G. G., & Hippert, H. S. (2020). Short-term Load Forecasting using Combined Data from Several Weather Stations. *International Journal of Advanced Engineering Research and Science*, *7*(9), 318–328. https://doi.org/10.22161/ijaers.79.38

Novak, D. R., Bailey, C., Brill, K. F., Burke, P., Hogsett, W. A., Rausch, R., & Schichtel, M. (2014). Precipitation and temperature forecast performance at the weather prediction center. *Weather and Forecasting*, *29*(3), 489–504. https://doi.org/10.1175/WAF-D-13-00066.1

Palmer, T. N., & Tibaldi, S. (1988). On the prediction of forecast skill. *Monthly Weather Review*, *116*(12), 2453–2480. https://doi.org/10.1175/1520-0493(1988)116<2453:OTPOFS>2.0.CO;2

Papalexopoulos, A. D., & Hesterberg, T. C. (1990). A regression-based approach to short-term system load forecasting. *IEEE Transactions on Power Systems*, *5*(4), 1535–1547. https://doi.org/10.1109/59.99410

Park, D., Mohammed, O., Azeem, A., Merchant, R., Dinh, T., Tong, C., Farah, J., & Drake, C. (1993). Load curve shaping using neural networks. *Proceedings of the 2nd International Forum on Applications of Neural Networks to Power Systems, ANNPS 1993*, 290–295. https://doi.org/10.1109/ANN.1993.264332

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A.,

Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*. https://doi.org/10.48550/arXiv.1201.0490

Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., & Thuerey, N. (2020). WeatherBench: A Benchmark Data Set for Data-Driven Weather Forecasting. *Journal of Advances in Modeling Earth Systems*, *12*(11). https://doi.org/10.1029/2020MS002203

Sandels, C., Widén, J., Nordström, L., & Andersson, E. (2015). Day-ahead predictions of electricity consumption in a Swedish office building from weather, occupancy, and temporal data. *Energy and Buildings*, *108*, 279–290. https://doi.org/10.1016/j.enbuild.2015.08.052

Sangamwar, S. (2019). Grouping Calendar Variables for Electric Load Forecasting [The University of North Carolina at Charlotte]. In *ProQuest Dissertations and Theses*. https://librarylink.uncc.edu/login?url=https://www.proquest.com/docview/2210140003?accountid=14605

Santiago, I., Moreno-Munoz, A., Quintero-Jiménez, P., Garcia-Torres, F., & Gonzalez-Redondo, M. J. (2021). Electricity demand during pandemic times: The case of the COVID-19 in Spain. *Energy Policy*, *148*, *A*. https://doi.org/10.1016/j.enpol.2020.111964

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and Statistical Modeling with Python. *Proceedings of the 9th Python in Science Conference*. http://statsmodels.sourceforge.net/

Segarra, E. L., Du, H., Ruiz, G. R., & Bandera, C. F. (2019). Methodology for the Quantification of the Impact of Weather Forecasts in Predictive Simulation Models. *Energies*, *12*(7), 1309. https://doi.org/10.3390/EN12071309

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

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

Stern, H., & Davidson, N. E. (2015). Trends in the skill of weather prediction at lead times of 1-14 days. *Quarterly Journal of the Royal Meteorological Society*, *141*(692), 2726–2736. https://doi.org/10.1002/qj.2559

Taieb, S. Ben, & 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

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

Taylor, J. W., & Buizza, R. (2002). Neural network load forecasting with weather ensemble predictions. *IEEE Transactions on Power Systems*, *17*(3), 626–632. https://doi.org/10.1109/TPWRS.2002.800906

Taylor, J. W., & McSharry, P. E. (2007). Short-term load forecasting methods: An evaluation based on European data. *IEEE Transactions on Power Systems*, *22*(4), 2213–2219. https://doi.org/10.1109/TPWRS.2007.907583

Teisberg, T. J., Weiher, R. F., & Khotanzad, A. (2005). The economic value of temperature forecasts in electricity generation. *Bulletin of the American Meteorological Society*, *86*(12), 1765–1772. https://doi.org/10.1175/BAMS-86-12-1765

Thornes, J. E., & Stephenson, D. B. (2001). How to judge the quality and value of weather forecast products. *Meteorological Applications*, *8*(3), 307–314. https://doi.org/10.1017/S1350482701003061

Tudose, A. M., Picioroaga, I. I., Sidea, D. O., Bulac, C., & Boicea, V. A. (2021). Short-Term Load Forecasting Using Convolutional Neural Networks in COVID-19 Context: The Romanian Case Study. *Energies*, *14*(13), 4046. https://doi.org/10.3390/en14134046

Vallance, L., Charbonnier, B., Paul, N., Dubost, S., & Blanc, P. (2017). Towards a standardized procedure to assess solar forecast accuracy: A new ramp and time alignment metric. *Solar Energy*, *150*, 408–422. https://doi.org/10.1016/j.solener.2017.04.064

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

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

Wang, Z., Hong, T., & Piette, M. A. (2020). Building thermal load prediction through shallow machine learning and deep learning. *Applied Energy*, *263*, 114683. https://doi.org/10.1016/J.APENERGY.2020.114683

Wernli, H., Hofmann, C., & Zimmer, M. (2009). Spatial forecast verification methods intercomparison project: Application of the SAL technique. *Weather and Forecasting*, *24*(6), 1472–1484. https://doi.org/10.1175/2009WAF2222271.1

Weron, R. (2006). *Modeling and forecasting electricity loads and prices: A statistical*

approach. John Wiley & Sons. https://doi.org/10.1002/9781118673362

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

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

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

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

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

Yang, D., Alessandrini, S., Antonanzas, J., Antonanzas-Torres, F., Badescu, V., Beyer, H. G., Blaga, R., Boland, J., Bright, J. M., Coimbra, C. F. M., David, M., Frimane, Â., Gueymard, C. A., Hong, T., Kay, M. J., Killinger, S., Kleissl, J., Lauret, P., Lorenz, E., … Zhang, J. (2020). Verification of deterministic solar forecasts. *Solar Energy*, *210*, 20–37. https://doi.org/10.1016/j.solener.2020.04.019

Yang, D., Wang, W., & Hong, T. (2022). A historical weather forecast dataset from the European Centre for Medium-Range Weather Forecasts (ECMWF) for energy forecasting. *Solar Energy*, *232*, 263–274. https://doi.org/10.1016/J.SOLENER.2021.12.011

Zhang, F., Qiang Sun, Y., Magnusson, L., Buizza, R., Lin, S. J., Chen, J. H., & Emanuel, K. (2019). What is the predictability limit of midlatitude weather? *Journal of the Atmospheric Sciences*, *76*(4), 1077–1091. https://doi.org/10.1175/JAS-D-18-0269.1

Zhao, J., Duan, Y., & Liu, X. (2018). Uncertainty Analysis of Weather Forecast Data for Cooling Load Forecasting Based on the Monte Carlo Method. *Energies*, *11*(7), 1900. https://doi.org/10.3390/EN11071900

Zhao, J., & Liu, X. (2018). A hybrid method of dynamic cooling and heating load forecasting for office buildings based on artificial intelligence and regression analysis. *Energy and Buildings*, *174*, 293–308. https://doi.org/10.1016/J.ENBUILD.2018.06.050

Zhou, Q., Wang, S., Xu, X., & Xiao, F. (2008). A grey-box model of next-day building thermal load prediction for energy-efficient control. *International Journal of Energy*

Research, *32*(15), 1418–1431. https://doi.org/10.1002/ER.1458

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