Quantifying Forecast Uncertainty in the Energy Domain

Mohammad Saber
Marquette University

QUANTIFYING FORECAST UNCERTAINTY IN THE ENERGY DOMAIN

by

Mohammad Saber, B.S., M.S.

A Dissertation submitted to the Faculty of the Graduate School,
Marquette University,
in Partial Fulfillment of the Requirements for
the Degree of Doctor of Philosophy

Milwaukee, Wisconsin

December 2017

ABSTRACT

QUANTIFYING FORECAST UNCERTAINTY IN THE ENERGY DOMAIN

Mohammad Saber, B.S., M.S.

Marquette University, 2017

This dissertation focuses on quantifying forecast uncertainties in the energy domain, especially for the electricity and natural gas industry. Accurate forecasts help the energy industry minimize their production costs. However, inaccurate weather forecasts, unusual human behavior, sudden changes in economic conditions, unpredictable availability of renewable sources (wind and solar), etc., represent uncertainties in the energy demand-supply chain. In the current smart grid era, total electricity demand from non-renewable sources influences by the uncertainty of the renewable sources. Thus, quantifying forecast uncertainty has become important to improve the quality of forecasts and decision making.

In the natural gas industry, the task of the gas controllers is to guide the hourly natural gas flow in such a way that it remains within a certain daily maximum and minimum flow limits to avoid penalties. Due to inherent uncertainties in the natural gas forecasts, setting such maximum and minimum flow limits a day or more in advance is difficult. Probabilistic forecasts (cumulative distribution functions), which quantify forecast uncertainty, are a useful tool to guide gas controllers to make such tough decisions.

Three methods (parametric, semi-parametric, and non-parametric) are presented in this dissertation to generate 168-hour horizon probabilistic forecasts for two real utilities (electricity and natural gas) in the US. Probabilistic forecasting is used as a tool to solve a real-life problem in the natural gas industry. A benchmark was created based on the existing solution, which assumes forecast error is normal. Two new probabilistic forecasting methods are implemented in this work without the normality assumption.

There is no single popular evaluation technique available to assess probabilistic forecasts, which is one reason for people's lack of interest in using probabilistic forecasts. Existing scoring rules are complicated, dataset dependent, and provide less emphasis on reliability (empirical distribution matches with observed distribution) than sharpness (the smallest distance between any two quantiles of a CDF). A graphical way to evaluate probabilistic forecasts along with two new scoring rules are offered in this work. The non-parametric and semi-parametric probabilistic forecasting methods outperformed the benchmark method during unusual days (difficult days to forecast) as well as on other days.

ACKNOWLEDGEMENTS

Mohammad Saber, B.S., M.S.

This work would not have been possible without the help of the financial support from the Electrical and Computer Engineering (EECE) department and the GasDay™ laboratory at Marquette University, and assistance of my doctoral advisor Dr. Richard Povinelli.

I would like to thank Drs. Richard Povinelli, George Corliss, and Ronald Brown for their thoughtful feedback on my work from the early stage of my research. Suggestions I have received during last four years are not only limited to academic. The weekly GasDay and KID seminars were helpful to get continuous feedback on my work. Also, I must thank my committee members Drs. Farrokh Nourzad, Henry Medeiros, and Ting Lin for their quick response and valuable ideas to improve this dissertation.

I would like to express my gratitude to GasDay laboratory for not only financial support, but also providing a great learning environment with fast computational resources. Special thanks to Thomas Quinn, for not only continuous funding but also arranging several seminars with GasDay customers, which added extra dimension to my research. Many thanks to Catherine Porter, she is like a mother in the GasDay lab, and cheering up everyone when someone looks depressed. Thanks to GasDay laboratory graduate students for all their remarks, ideas, constructive criticism, and support.

I am grateful to Tom Connery and Tyler Stephens, GasDay customers who have introduced some of the challenges for gas controllers in several GasDay seminars. The initial idea and motivation for this research work came from trying to solve one of the challenges mentioned in their presentation.

I would like to dedicate this work to my parents Shahrya Quadir and Mohammad Nurul Quadir, my wife Sumaiya Ahsan, my elder brother Mohammad Shoaib, extended family members, and friends. Their sacrifice, patience, love, and continuous mental support work as huge inspiration to overcome all roadblocks in my way to graduation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION TO FORECAST UNCERTAINTY QUANTIFICATION

This dissertation focuses on improving the quality of natural gas and electric demand forecasts by studying two types of forecasts: point forecasts and probabilistic forecasts. A point forecast provides a single-valued best estimate, whereas a probabilistic forecast helps to quantify the uncertainty of a future event with a cumulative distribution function (CDF). Probabilistic forecasts add useful information to point forecasts, when the forecasted event is associated with uncertain factors. For example, energy demand is highly correlated with weather conditions, human behavior, and economic conditions, which are all uncertain. In this dissertation, probabilistic forecasts are used to quantify the forecast uncertainties in the energy domain. Section 1.1 describes the problem statement of this dissertation. A concise overview of the electricity and natural gas industries are included in Section 1.2. The importance of quantifying forecast uncertainty in the energy industry and the contributions of this dissertation are explained in the following sections. Finally, the outline of the dissertation is given at the end of this chapter.

1.1 Problem Statement: Quantifying Forecast Uncertainty

The problem addressed by this dissertation has two parts. The first part describes the problem statement in the context of the natural gas industry, and the second part provides the electricity industry version of the problem statement.

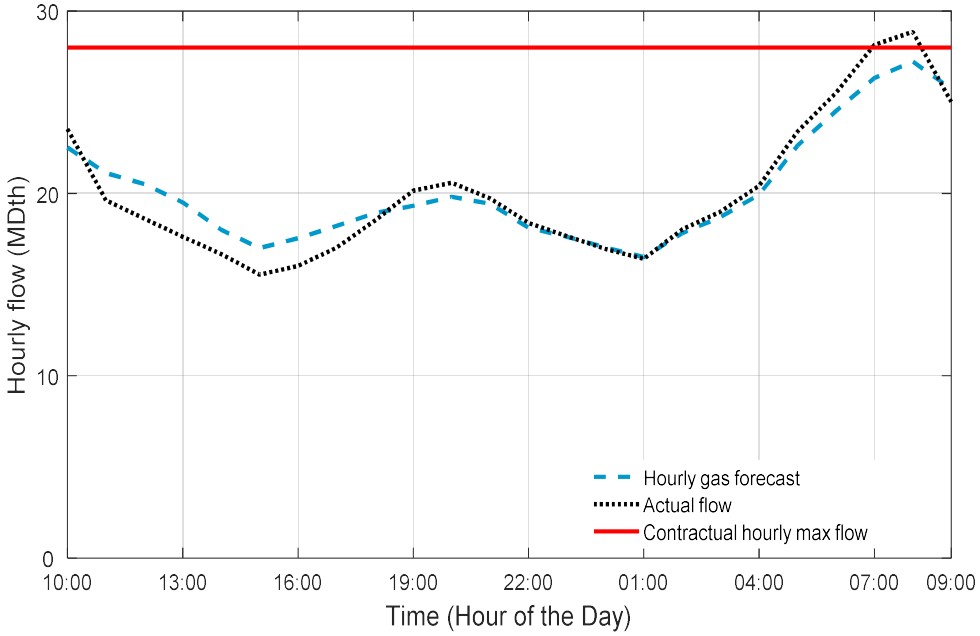

Figure 1.1: Sample hourly actual and forecasted gas flow.

In the natural gas industry, gas controllers continuously monitor the contractual hourly maximum flow, as shown in Figure 1.1. This is the hourly maximum gas draw limit from a gas supply pipeline to a gas utility. This limit is usually set in the morning for the coming gas day (9 A.M. – 9 A.M., Chicago time). If the gas demand for any hour crosses this limit, then the gas utility has to purchase extra gas from the spot market, which is sometimes several times higher than the usual price [1]. Currently, gas utilities use point forecasts to decide the level of the contractual hourly maximum flow. However, point forecasts do not convey the uncertainty associated with a future event. It is expected that the actual flow will differ from the point forecasts most of the time. In Figure 1.1 (gas flow at 10 A.M. means the total gas flow from 9 A.M. to 10 A.M.), a cartoon scenario is shown in which point forecasts do not provide enough information about the possibility of crossing the maximum gas flow limit. Probabilistic forecasts, which

provide a more complete view of a future event, are useful tools for the gas controllers to determine the risk of exceeding the contractual hourly maximum flow.

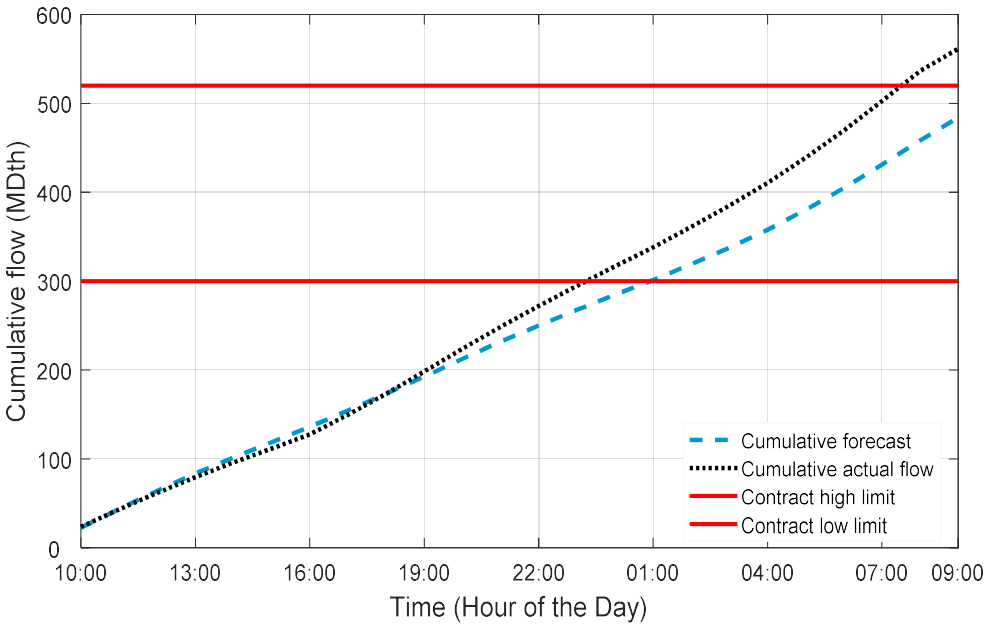

Figure 1.2: Sample hourly cumulative actual and forecasted gas flow.

Similarly, the gas controllers have to guide the gas flow such that at the end of the gas day, the cumulative gas flow for the gas day remains within a certain range. The upper flow limit is called the contract high limit, and the lower limit is called the contract low limit. Both are illustrated in Figure 1.2. These limits are set when the gas purchasing contract was signed between the gas utilities and the gas suppliers early in the morning (before the start of the gas day). Consequences of crossing the boundaries are similar to penalties for crossing the contractual hourly maximum flow. These problems were brought to our attention by GasDay^{TM} customers Tom Connery and Tyler Stephens in describing some of the challenges of their operations as gas controllers in a local

distribution company (LDC) [2]. The point forecasts cannot provide enough information to help the gas controller to guide the cumulative actual flow within the contract high and low limit. On the other hand, the probabilistic forecasts offer additional guidance to support crucial decisions.

In the electricity industry, forecasts are required to make planning and maintenance decisions [3, 4, 5]. One of the major differences between natural gas and electricity production is that the electricity plants produce electricity when it is required and supply it immediately to the customers because storage of electricity is expensive. Hence, any excess production of electricity means wasted energy, and a shortage of electricity may lead to a blackout. To avoid a blackout in case of higher than expected demand, electricity distribution companies can buy extra electricity on the spot market (if available), which usually is higher than the normal production costs. Thus, electricity distribution companies use electricity demand forecasts for careful planning of electricity purchases and generation.

In the smart grid era [6], electricity demand has become highly unpredictable [6, 7, 8]. The smart grid is a two-way communication between electricity consumers and producers. All types of power plants (renewable, oil, coal, gas, and nuclear) are connected to a single (smart) grid to provide customers an opportunity to choose the best energy options for power. The U.S. government is encouraging electricity consumers to use clean energy sources by providing incentives [9] to meet the goal of reducing carbon emission by 26-28% below 2005 levels by 2025 [10]. Therefore, the renewable sources such as wind and solar are becoming major contributors to the smart grid system.

However, wind and solar power are not available all the time, and the intensity of these natural sources vary. If renewable sources are not available, other sources (oil, gas, coal, nuclear) are needed to fill the gap to insure an uninterrupted power supply. When renewable sources are abundant, the demand for nonrenewable source generated electricity is expected to be less, due to subsidies on renewable sources. The uncertainty of the availability of wind and solar energy affect the predictability of nonrenewable sources. Existing point forecasting methods do not work well to predict energy demand in the smart grid era. Probabilistic forecasts, which can quantify the forecast uncertainties, are considered as a better option for forecasting energy demand [11].

The next section provides a brief overview of the energy industry in the U.S. The following sections describe the importance of probabilistic forecasts, contributions made in this dissertation, and an outline of this dissertation.

1.2 Energy Industry Overview

Petroleum, natural gas, coal, renewable energy (wind, solar, biofuels, wood, hydro, geothermal, and biomass waste), and nuclear electric power are the primary sources of energy consumption in the U.S. [12]. Electricity, which is considered as a secondary source of energy, can be generated from all primary sources mentioned above. Datasets used in this dissertation to verify the effectiveness of our proposed models are collected from a local natural gas distribution company and a local electricity distribution

company in the U.S. This section surveys both natural gas and electricity industries in short for a better understanding of the methods proposed in this dissertation.

1.2.1 Natural Gas Industry

The natural gas supply chain has three major components: natural gas production, transportation, and supply to the end user (Figure 1.3). Natural gas is produced from oil and gas wells and stored in a gas plant storage. Pipeline companies transport the natural gas to local distribution companies (LDCs). Finally, LDCs supply natural gas to the end users.

LDCs are responsible for ensuring an uninterrupted gas supply. This is especially important during the heating season, typically November to March. There are four categories of end users who are considered as "customers" of the LDCs.

Residential customers use natural gas for cooking, heating and cooling spaces, and running appliances such as cloth dryers, pool and Jacuzzi heaters, fireplaces, barbecues, and outdoor lights [13].

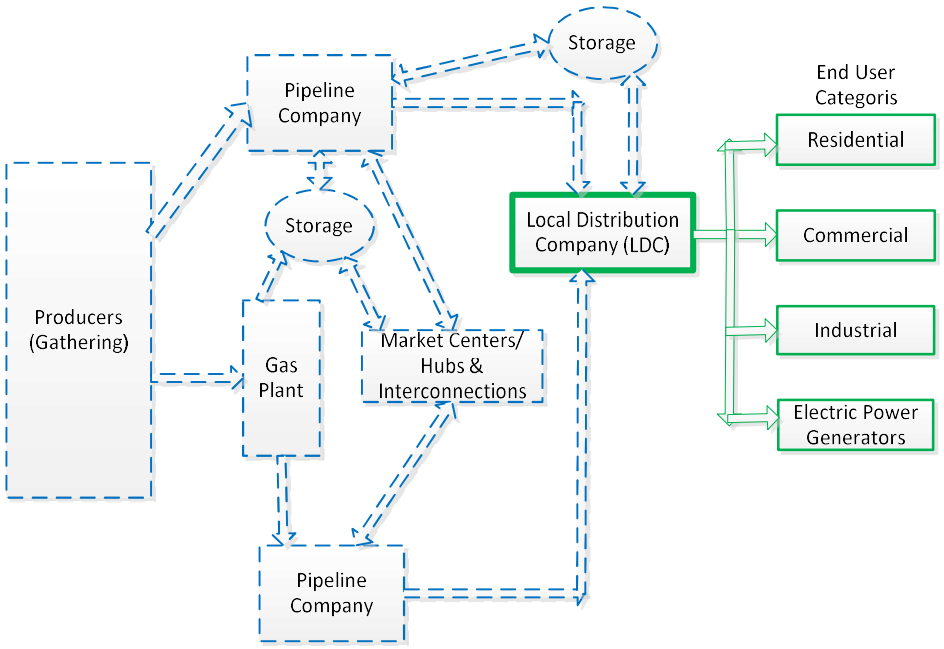

Figure 1.3: Natural gas distribution segments (adapted from [14]).

Commercial customers use natural gas for space heating and cooling, and steam heating of commercial buildings such as office buildings, schools, churches, hotels, restaurants, and government buildings. Members of the food service industry use natural gas for commercial cooking [15].

Both residential and commercial use of natural gas is weather dependent. In the winter, natural gas use peaks due to space heating, and in the summer, it is sometimes used for cooling.

Industrial customers consume natural gas for the same purposes as commercial and residential customers such as heating, cooling, and cooking [16]. Natural gas is used also in industry for metal preheating, drying, dehumidification, glass melting,

incineration, waste treatment, and food processing. It also is used as a feedstock in the plastic, fertilizer, anti-freeze, and fabrics industries.

Electric power generators use natural gas for generating electricity. Currently, natural gas is one of the most popular fuels for electricity generation as an environmentally friendly and low-cost source. In 2016, the U.S. Energy Information Administration (EIA) published the last 15 years (2000-2015) of total energy production in the U.S., where the trend suggests that the use of natural gas will increase further into the future [17].

1.2.2 Electricity Industry

Electricity is generated from the conversion of another energy source such as coal, solar, wind, natural gas, oil, water, or nuclear. Figure 1.4 illustrates the supply chain of electricity. Step-up transformers are used to transfer electricity from power plants to local distribution areas. Then step-down transformers are used to transfer electricity into the distribution lines. Finally, the customers receive electricity from distribution lines.

Electricity usage in the United States can be divided into four major sectors [18]:

The industrial sector is the largest consumer of electricity. This sector uses electricity for manufacturing, agriculture, mining, and construction.

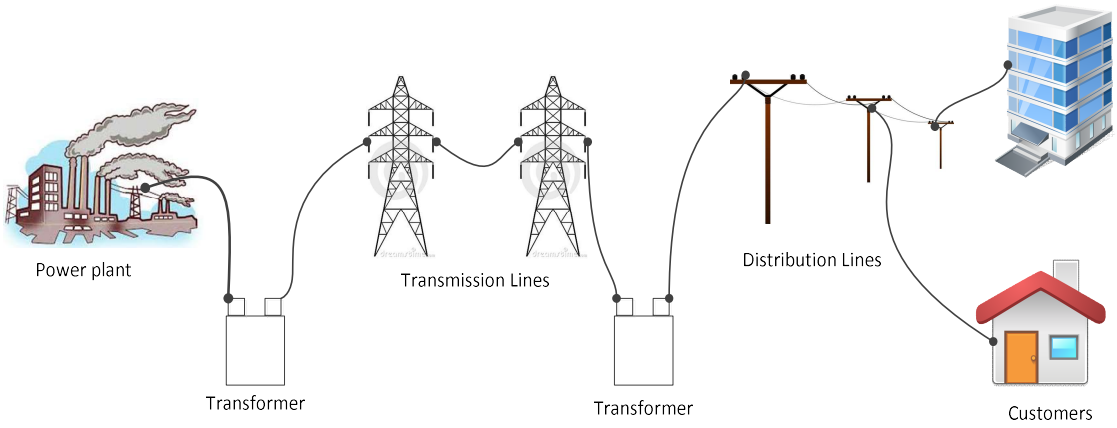

Figure 1.4: Electricity supply chain (adapted from [19]).

The transportation section is the second largest customer of electricity. This sector is growing fast due to popularization of electric cars. In the transportation sector, electricity is used also in trucks, buses, motorcycles, trains, aircraft, boats, ships, and barges. However, the use of electricity by trains, aircrafts, boats, ships, and barges are not related to daily load forecasting in the U.S.

The residential sector uses electricity for lighting, heating and cooling, cooking, charging electrical equipment, and entertainment.

The commercial sector includes office buildings, schools, universities, religious places, apartments, hospitals, warehouses, hotels, restaurants, and shopping malls. This sector uses electricity for heating, lighting, cooling, and powering electric equipment.

Residential and commercial customers' electricity use is weather dependent. In the summer, the electricity is used for cooling; and in winter, electricity is used for heating.

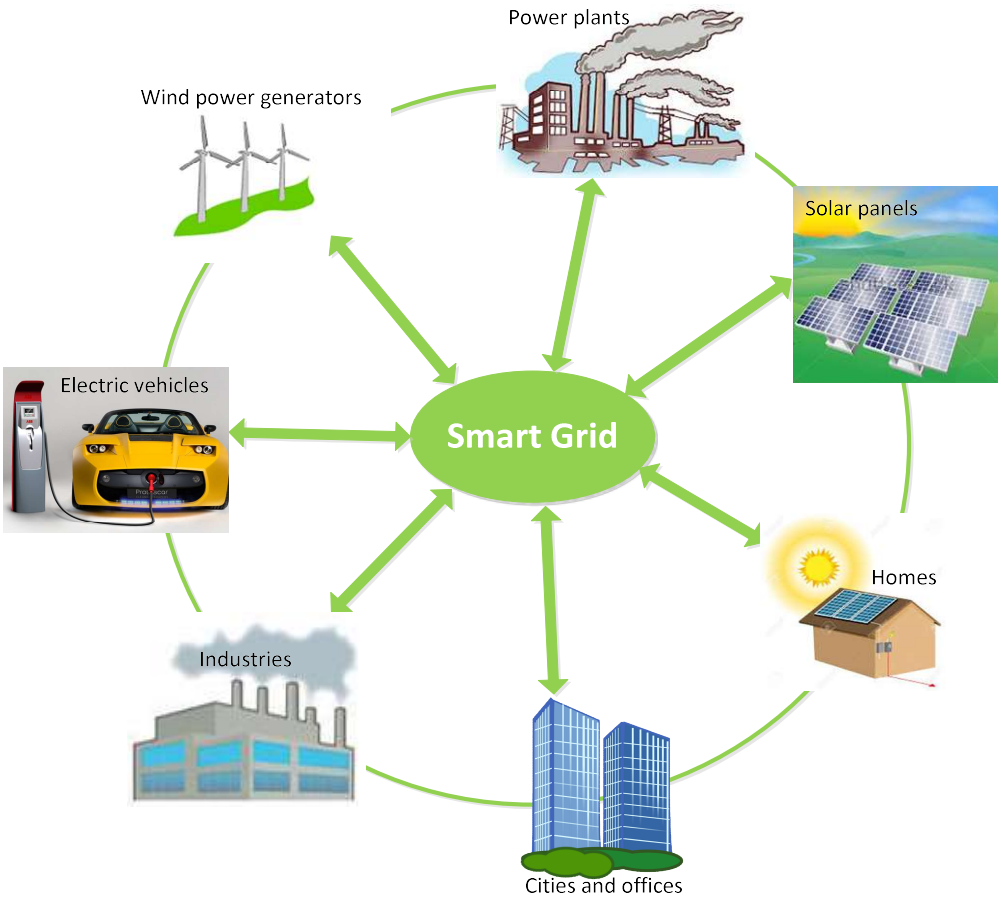

Figure 1.5: An overview of the smart grid (similar to [20, 21]).

The current electricity energy supply chain is shown in Figure 1.4 has one-way communication from the producer of electricity (power plants) to the customers. The fault tolerance of the current interconnected grid system is not robust. A single point of failure can interrupt the power supply of other grids connected to that point [22]. The process of identifying the problem area and reestablishing the connection is manual and time consuming. Hence, in 2007, the U.S. congress passed a law mandating modernization of the electric grid [23]. The new electricity power grid, which is able to

find the disconnected region automatically and reroute the electricity supply to ensure uninterrupted power supply, is known as the “smart grid” [24].

The smart grid provides bidirectional communication between customers and producers (power plants) [25]. Figure 1.5 illustrates the important components of the smart grid. The use of the smart meters (an electric device that records electric energy at shorter-than-hourly intervals and returns the result back to utilities and customers at least daily [26]) allows customers to check their electricity consumption in near real time. Hence, customers can set their smart appliances to work during off-peak hours to reduce their electricity bills [27]. In the new grid system, residential and commercial customers can act as energy producers through rooftop solar panels. Customers can sell the excess electricity to local utility companies. Residential customers can store electricity in their electric vehicle or in a home power station (storage connected to solar generation) during off-peak hours and use it during peak hours. In the current energy supply model, electricity storage is not considered due to its high cost. However, the electric car may play a major role of storing electricity in the future. Renewable energy sources such as wind and solar energy are important parts of the smart grid. Recently, the U.S. government set a target to reduce climate pollution by 26-28 percent below 2005 levels by 2025 [10]. Thus, it is expected that renewable power plants will grow rapidly in the near future with the help of subsidies provided by the U.S. government [9].

The next section will explain the importance of probabilistic forecasts in the energy industry.

1.3 Importance of Probabilistic Forecasts

In both the natural gas and electricity industries, forecasts are required every day to plan for the needed energy supply [28]. An accurate point forecast leads to efficient planning, which maximizes the profit of doing business for local energy distribution companies by reducing the production and distribution cost of natural gas and electricity. Over- and under-forecasts both increase the unit price of natural gas (Figure 1.2) and electricity distribution because of last minute adjustments. Perfectly accurate forecasts are not possible due to inherent uncertainties involved with the output of mathematical models. Thus, understanding the uncertainties in natural gas and electricity forecasting models allow energy distributors to make more cost-effective decisions.

Uncertainties are classified into two categories: “aleatory” (irreducible or random) and “epistemic” (reducible, but enough information is not available at this moment) [29]. In natural gas and electricity forecasting, uncertainties arise from different sources. Uncertain weather forecasts (temperature, wind speed, dew point), which are used as inputs to energy forecasting, may cause uncertainty in the model. The impact of random human behavior on energy consumption cannot be measured accurately. Numerical errors (rounding error, discretization errors) may cause epistemic uncertainty. Good data helps generate better mathematical models with less uncertainty; however, in practice, it is impossible to obtain perfect datasets. Forecasting models contain unknown errors if all the required factors are not included in the model. On the other hand, if all known factors are included, the model may become unpractical to use because of a huge calculation

burden and overfitting. In addition, all the factors influencing a forecasting model are not known [30]. Thus, uncertainties cannot be eliminated completely from forecasting models. Hence, probabilistic forecasts, which consider forecast uncertainties, are better tools than the point forecast for making business decisions (see Figures 1.1 and 1.2).

1.4 Contributions of Quantifying Forecast Uncertainty

This dissertation makes three major contributions. First, it provides several ways to generate probabilistic forecasts by analyzing error distributions from an existing point forecasting model.

Secondly, there are several metrics for scoring probabilistic forecasting methods. However, existing techniques are dataset dependent, so the scores using the same evaluation metric in different research articles are not comparable. Thus, a new evaluation metric (scoring rule) has been implemented that is not dataset dependent.

Finally, probabilistic forecasts are used rarely in natural gas forecasting, and they are new in electricity load forecasting. Therefore, showing the applications of probabilistic forecasts for efficient decision-making in the energy industry (natural gas and electricity) will encourage future researchers to work more in this field.

1.5 Outline of the Dissertation

Chapter 2 provides a literature review of existing methods used for point forecasts, probabilistic forecasts, and evaluation metrics to assess probabilistic forecasts. Both point forecasts, and probabilistic forecasts are studied in three subsections: survey papers, statistical approaches, and machine learning approaches.

Chapter 3 contains proposed methods of generating probabilistic forecasts. This chapter explains the point forecasting methods used in this research to generate probabilistic forecasts. The second part of the chapter presents two new evaluation metrics (scoring rules), named Quantile Calibration Score (QCS) and Percentage Quantile Calibration Score (PQCS), respectively, for measuring the goodness of probabilistic forecasts.

The first part of Chapter 4 describes data collection and processing for implementation of the three probabilistic forecasting methods including a benchmark. The second part of the chapter compares the performance of probabilistic forecasting methods with a naive benchmark model on real data collected from natural gas utilities and electricity utilities. All probabilistic forecasting methods are evaluated using the scoring rules presented in Chapter 3 and two widely used existing scoring rules. Performance analysis of the probabilistic forecasting methods during unusual days (such as bitter cold days, big temperature swings) are included also in this chapter.

Chapter 5 concludes the dissertation with a summary, conclusions, and suggestions for future work.

CHAPTER 2

QUANTIFYING FORECAST UNCERTAINTY: LITERATURE REVIEW

Energy demand forecasting is more than a century old problem [31]. A good demand forecast saves millions of dollars and ensures quality of service in the energy industry [1, 32, 33, 34]. Pierre Pinson, in his Ph.D. dissertation, established that advanced point forecasts can help to reduce regulation cost by 38%, and knowing reliably the uncertainty can decrease the regulation cost by another 39% [35, 36]. Although quantifying forecast uncertainty may support better decisions in the energy industry, few research articles have been published on quantifying forecast uncertainty compared to thousands of journal articles on point forecasting [37]. This chapter presents existing methods for generating point forecasts and probabilistic forecasts. Existing evaluation techniques of probabilistic forecasts are reviewed at the end of the chapter.

2.1 Point Forecasts

Forecasts has been used in economics, meteorology, and energy for a long time [38, 39, 40]. In 1980, the IEEE Load Forecasting Working Group published a bibliography of load forecasting papers; where the earliest article on point forecasting was published in 1918 [38]. For the last three decades, thousands of papers have been published on point forecasting techniques [37]. It is difficult to fit every articles in load forecasting field in one chapter, because there are so many. Thus, survey papers are

useful to cover many articles concisely. The next subsection contains a review of highly cited papers on load forecasting. The subsequent subsections review load forecasting literature in two parts: statistical approaches and machine learning approaches. Sometimes it is difficult to separate statistical methods from machine learning approaches. However, this division is useful to help understand the literature. The main focus of this dissertation is to quantify uncertainty using probabilistic forecasts. However, initially a point forecasting method was required to generate the desired probabilistic forecasts. This section provides the background needed to understand the point forecasting methods used in this dissertation.

2.1.1 Point Forecasting Review

Gross et al. (1987) offered a tutorial review of short-term load forecasting (STLF) by discussing: 1) importance and application of STLF; 2) essential factors to consider during load forecasting; 3) different forecasting models of STLF; 4) practical considerations to implement STLF; and 5) use of STLF in control center environments [4]. The authors mention weather, time, economic, and random effects as four major driving factors of STLF. Three types of performance measures are proposed in that article: a) accuracy, b) ease of use, and c) outlier detection and correction capabilities. The authors point out many important issues such as error analysis, holiday effects, bad data handling, and complete automation of the forecasting process. These issues are still considered as major concerns for STLF.

Moghram et al. (1989) did a comparative evaluation of five short-term load forecasting techniques: 1) multiple linear regression; 2) stochastic time series; 3) general exponential smoothing; 4) state space method; and 5) knowledge-based expert system [41]. Data from a utility in the southern U.S. was used to compare these five forecasting methods. The knowledge-based expert system outperformed the other four approaches. However, the authors made several unreasonable assumptions. For example, perfect weather forecasts are assumed, and the base loads of weekends and weekdays are assumed to be the same. The proposed point forecasting method used in this dissertation does not require any of these assumptions.

Bunn (2000) provided a review of some important issues for making price-sensitive decisions in the competitive electricity market [42], such as the effect of forecasting error on profit, importance of day-ahead weather based forecasts, dynamic price response, increase of distributed and embedded electricity generation, and the importance of market share forecasting. Specific emphasis has been given to segmentation of forecast variables, combination of the forecasting methods, and the use of neural networks for electricity load and price forecasting. The author suggest combining traditional time series methods with artificial neural networks (ANNs).

Hippert et al. (2001) reviewed a collection of ANN-based STLF articles from 1991 to 1999 [43]. They described the design process of STLF in four steps: 1) data pre-processing; 2) ANN design; 3) ANN implementation; and 4) validation. The authors identified two major concerns about ANN-based methods: over-parameterization and non-systematic testing. The over-parameterization leads to “overfitting” the data, which

gives better training results, but worse performance on unseen data. According to the authors' observations, some of the work on ANNs did not follow standard statistical reporting procedures. Additionally, results were not compared with standard benchmarks.

Alfares et al. (2002) presented a review by classifying electric load forecasting into nine categories: 1) multiple regression; 2) exponential smoothing; 3) iterative reweighted least-squares; 4) adaptive load forecasting; 5) stochastic time series; 6) auto regressive moving average models with exogenous inputs (ARMAX) models based on genetic algorithms; 7) fuzzy logic; 8) artificial neural networks (ANN); and 9) knowledge based expert systems [44]. The paper briefly described the methods and offered advantages and disadvantages. The authors mentioned fuzzy logic, genetic algorithms, expert systems, and neural networks as the most popular techniques among the nine categories based on number of publications in load forecasting in the early 2000s.

Weron (2006) offered a comprehensive review of statistical tools that can be used to analyze and forecast electricity load and price [3], such as seasonal decomposition, exponential smoothing, spike preprocessing, regime-switching models, and jump-diffusion models. A detailed structure of the electricity market in Europe, North America, Australia, and New Zealand was provided. Sixteen cases studies were included. Implementation of different statistical techniques, electricity load and price data, and learning toolboxes in MATLAB and SAS were provided.

Hong (2010) reviewed the last four decades of load forecasting articles in his Ph.D. dissertation [45]. He implemented three techniques to generate point forecasts for a

medium size utility in the U.S.: 1) multiple linear regression, 2) fuzzy regression, and 3) artificial neural networks (ANN). A brief tutorial on multiple linear regression, polynomial regression, interaction effects, fuzzy logic, artificial neural networks, most useful lag terms, weekend effects, and holiday effect are provided. A benchmark for each of the methods was created for comparison. In the case study, linear regression outperformed the other two methods. However, it cannot be concluded that linear regression always outperforms fuzzy logic and neural networks. The outcome may be different in different case studies [46].

In 2012, the IEEE Working Group on Energy Forecasting organized a Global Energy Forecasting Competition (GEFCom2012) to bring together state-of-art methods for energy forecasting [47]. The competition attracted hundreds of participants, who contributed many novel ideas in two tracks: 1) hierarchical load forecasting and 2) wind power forecasting. Hong et al. (2014) reviewed the top ten winning methods from both tracks [47]. In the load forecasting track, four teams used multiple linear regression, two of them applied gradient boosting machines, and the rest of the teams used random forecasts, generalized additive model, wavelet decomposition, and neural networks. Some of the teams performed additional tasks, such as data cleaning, combining forecasts of more than one method, and modeling holidays. A follow-up competition of GEFCom2012 was held in 2014 focusing on probabilistic forecasts, which is introduced in Section 2.2

Load forecasting papers can be divided into two broad categories based on implementation approaches: statistical and machine learning, which are presented in the next two sections.

2.1.2 Statistical Approaches for Forecasting

Linear regression is one of the simplest and most effective statistical approaches used for load forecasting. Amral et al. implemented three different multiple linear regression models to forecast 24-hour electricity load [48]. Data collected from Indonesia's South Sulewesi Power System was used as a case study. Separate datasets were used to forecast dry and rainy seasons, respectively. The authors mentioned that weather forecasting errors contribute heavily to the load forecasting error. However, forecasted weather data were not used in this experiment because it was not available. Thus, the presented load forecasting error is the best case scenario when the weather forecast is perfect. In this work, forecasts are made from both actual weather and forecasted weather datasets.

Hong et al. focused on benchmarking STLF [45, 49]. His paper proposed a naïve multiple linear regression (MLR) short term load forecasting model considering linear trend, calendar variables, relationship between temperature and load, and interaction effects. This model has been used as a benchmark in a U.S. utility since 2009 and applied in a Canadian utility for load forecasting. The benchmark model was used in [50] to compare with a long term load forecast. A simplified method adapted from [51] was used to produce a one-year ahead load forecast from the GEFComm2012 dataset [47]. The

main contribution of that paper was finding appropriate temperature lag terms. The energy use was highly correlated with temperature. However, energy users tend to react late during temperature changes, which is labelled as a “recency effect” (lag) in [50].

Xie et al. used the benchmark model presented in [45] to forecast load per customer [52]. The customer attrition rate was forecast using a generalized linear model (GLM) implemented in SAS [53]. A long-term energy forecast was calculated by multiplying per customer load by a projected number of customers. The MLR model proposed in [45] was used also as a benchmark in the hierarchical load forecasting track of GEFCom2012 [47].

One of the main challenges of hierarchical load forecasting is to deal with weather data from multiple weather stations for an operating area or zone. Hong et al. used a MLR method to rate and rank weather stations of a region [54]. Weather stations were combined based on their ranks. Finally, various combinations of the subset of weather stations were ranked using the same MLR method to select the most useful set of weather stations for that zone.

Autoregressive (AR) and moving average (MA) models were first formulated by Yule in 1927 [55]. The widely used time series forecasting model ARMA (autoregressive moving average) is the combination of AR and MA models [56]. In 1970, Box and Jenkins integrated the existing knowledge to develop a three stage iterative process for time series identification, estimation, and verification, which is known as autoregressive integrated moving average (ARIMA) [57]. Since 1970, the ARIMA and ARMA

approaches were frequently mentioned as important load forecasting methods [3, 4, 28, 33, 41, 44, 58, 59, 60, 61].

Huang et al. proposed a method based on ARMA with a non-Gaussian model to forecast one-day-ahead hourly load and one-week-ahead daily peak load [60]. The performance of the model was tested on the electricity demand data collected from the Taiwan Power Company. Load forecasts were compared with traditional ARMA and artificial neural networks (ANN). Both traditional ARMA and improved ARMA models performed better than the ANN.

Exponential smoothing is a technique used in forecasting in which more weight is given to recent data, and the weight decreases exponentially for older data [62]. Exponential smoothing was introduced independently by Brown in 1956 [63, 64] and Halt in 1957 [65]. In 1960, Winters experimented using Halt's method [66]. Although Halt's original paper was not published until 2004 [67], this method became known as the Halt-Winter method.

In 2006, the International Institute of Forecasters (IIF) published a review on the progress of time series forecasting between 1982 and 2005 [68]. The review discussed 940 papers on time series forecasting focusing on five topics: 1) point forecasting methods, 2) count data forecasting, 3) forecasting evaluation, 4) combining forecasts, and 5) density (probabilistic) forecasts. The exponential smoothing and different variants of the ARMA model are used by most of the researchers for time series forecasting. Non-linear model such as artificial neural networks (ANNs) are used also for time series

forecasting. A vision for the next 25 years of time series forecasting by the IIF is included.

Taylor et al. compared a day-ahead electricity demand forecasting performance of exponential smoothing with linear regression, neural networks, ARMA, and two benchmarks based on regression in 2006 [69]. Experiments were conducted on data from two different regions (Rio de Janeiro, and England and Wales). Exponential smoothing had better forecasts among the six methods.

Adya et al. developed a rule-based forecasting (RBF) expert system which combined four statistical methods: random walk, linear regression, Holt's exponential smoothing, and Brown's exponential smoothing [70]. A heuristic method was developed to automate the process of weight exploration for RBF. A comparison between random walk, equal weight RBF, dynamic weight RBF, and automated weight RBF was shown based on 732 forecasts. The automation technique reduced the overhead cost of using RBF, which was the main contribution of that paper.

This section has reviewed recently published point forecasting methods based on statistical approaches such as MLR, ARIMA, and exponential smoothing. Findings from most of the literature are based on specific case studies. The contradictory demand of one method outperforming another method proves that there is no absolute winner in this field of research. Moreover, the length of the forecasting dataset used as a case study for most of the papers cannot provide definite conclusions. Data preprocessing such as cleaning and weather station selection are shown useful to improve forecast accuracy.

2.1.3 Machine Learning Approaches for Forecasting

This section reviews machine learning approaches such as artificial neural networks, support vector machines, fuzzy inference systems, genetic algorithms, and gradient boosting machines.

Artificial neural networks (ANNs) have been used for STLF since the early 1990s [43, 45]. Jain et al. used a clustering algorithm with ANNs for producing day-ahead load forecasts [58]. A clustering technique was applied to find similar days based on daily average and peak loads. Unfortunately, the authors used only 13 weeks of data to compare clustering-based ANN with a simple ANN. This is too little data to capture the seasonal effects of electricity demand. Their results showed that the use of clustering techniques with ANN improved forecast performance.

Wang et al. combined ARMA with ANNs to forecast the daily load of the Jiangmen Power Company [71]. Linear and non-linear components of a short-term time series were forecasted separately by ARMA and ANN models, respectively. Then both forecasts were combined to generate daily load forecasts. The experiment showed better performance of the combined method when compared with individual methods (ARMA or ANN). However, the testing dataset was very small (only 31 days).

Siddique et al. incorporated machine learning techniques such as ANN and Regression Tree (RT) to learn domain knowledge of time series by input feature transformation [61]. The proposed method was applied to forecast daily natural gas and electricity demand from two locations in the United States. The process of finding the

best set of inputs (AR terms, day of the year, and temperature) and models (ARMA, RT, and ANN) was automated. An advanced version of Siddique's linear regression model [46] has been adapted in this dissertation to generate point forecasts (see section 3.1).

Osman et al. performed quarterly correlation analysis of weather inputs (temperature, dew point, wind speed, and humidity) with electricity load. They used data from the Egyptian United System to determine the most correlated weather inputs for each quarter of the year [72]. The selected set of inputs were fed into an ANN model to forecast 24-hour ahead electricity demand. The proposed method performed better (especially in the summer) compared with a benchmark model (traditional ANN), which considered temperature as the most correlated input with electricity demand for all seasons of the year.

Qingle et al. pointed out that the very short-term load forecasting error based on ANNs is larger near peak loads [73]. Rough set theory (a set of decision rules) was applied to adjust the ANN-generated forecasts. Test results show that the use of a rough set theory adjustment significantly improved forecast accuracy. However, the test set was only 24 hours.

Ramos et al. used multiple layer ANN and Halt-Winter's exponential smoothing methods to generate 24 hour ahead load forecast for a Portuguese utility [74]. K-means clustering was used to find four distinct load profiles. Four models were created based on workdays, weekends, government holidays, and time interval (15 minutes and 1 hour) of recorded electricity demand. The authors concluded that the ANN outperformed an

exponential smoothing method based on forecast accuracy of a 24-hour period (first day of August 2011).

Xin-hui et al. used a three-layered ANN, a four-layered ANN, and a four layered ANN with a genetic algorithm (GA) for short-term load forecasting [75]. These models were trained with and without weather inputs. The test result showed better performance using weather inputs. Also the four-layered ANN performed better than the three-layered ANN, although the four-layer ANN had fewer neurons than three-layered ANN. No information was provided about the dataset.

Sun et al. used an extended Kalman filter (EKF) based ANN to forecast 24-hour ahead electricity demand for ISO New England [76]. EKF is used as a learning algorithm to train the ANN by treating the weights as a state [77]. ISO New England has a large geographic area containing 23 substations within two zones. Zonal load was forecast using a decoupled EKF. The substations with load patterns similar to the zonal load were forecast by simply calculating the proportion of zonal load. The rest of the substation load was forecasted using ANNs. The experimental result showed that the decoupling technique saved significant training time.

MATLAB organized a webinar in 2016 to demonstrate existing useful MATLAB toolboxes for generating short-term and long-term electricity load and price forecasts [78]. Neural networks and regression tree methods were used to produce 24-hour ahead load and price forecasts from the ISO New England dataset [79]. A detailed report for each of the methods for both load and price forecasting tracks is available online [80, 81,

82]. The forecasting demo is reproducible because the code, datasets, and documentation are all available online. A basic framework of point forecasting has been provided in the code. The MATLAB electricity load forecasting model was used as a benchmark to improve the proposed point forecasting method used in this dissertation.

A support vector machine (SVM) is an effective classifier [83]. This method is the second most used machine learning technique after ANN for load forecasting. Mohandes used SVMs to forecast electricity demand for the Eastern Province in Saudi Arabia [84]. Six years of hourly electric load were used in this experiment. Two outages, a clear uptrend of energy demand, and a seasonal effect (load was higher in summer than winter) were found during data analysis. Data was preprocessed by removing the seasonal effect, the trend, and the outliers. An autoregressive (AR) model was developed to compare the forecasting performance of the SVM. The SVM outperformed the baseline AR model.

Shu et al. developed a hybrid model based on SVM and self-organized maps (SOM) to forecast electricity load [85]. A SOM network was used to cluster the data into several subsets including anomalous days and regular days. Then, a group of 24 SVMs were trained to forecast the next the 24 hours of energy demand. The proposed model was tested on the New York City ISO electricity load [86]. The proposed hybrid network was compared with a SVM and existing methods used by the New York ISO.

Jin et al. used a grey forecasting model (a time series forecasting model, including a group of differential equations) and SVM to forecast electricity demand [87]. Although the single grey forecasting model did not provide good forecasts, two grey forecasting

models were combined to improve the forecasting performance. The grey forecasting model was used with a SVM to create a hybrid load forecasting model. The authors showed that the combination of SVM and grey forecasting model reduced the forecasting error significantly. However, the test dataset is small (5 hours of forecasts with 30 minute intervals).

Escobar et al. compared two machine learning techniques: SVM and adaptive neuro-fuzzy inference system (ANFIS), based on short term load forecasting [88]. The ANFIS is the combined load forecasting method originated from ANN and fuzzy inference systems (FIS). The SVM performed better than the ANFIS according to one week of forecasting performance. A variation of SVM, called least squared SVM (LS-SVM) was used by Espinoza et al. for short-term load forecasting [59]. The performance of the method was compared with a linear model with the same variables.

Ding proposed a decision tree based method to forecast long term electricity load in a developing area [89]. Fourteen economic factors were used to construct the decision tree. The original decision tree algorithm, ID3 [90] was modified by pruning (to avoid overfitting) and giving more weight to recent data. The improved decision tree method was used to forecast three years of load demand. The mean absolute percentage error (MAPE) of the forecast was used to compare ID3, MLR, cubic polynomial, exponential curve, compertz curve, and grey model. The decision tree method produced better forecasts for two out of three years.

Hamid et al. proposed an artificial immune system (AIS) learning algorithm as an alternative learning algorithm to train an ANN for short-term load forecasting [91]. The AIS learning algorithm has four steps: initialization, cloning, mutation, and feedforward [92]. The weight was updated based on the clonal selection theory [93]. Two sets of electricity demand data, one from Kuala Lumpur, Malaysia, and another from North Carolina, US, were used as case studies. The accuracy of the forecast measured in mean absolute percentage error (MAPE) and computational speed (number of iteration and time in seconds) was compared with an ANN trained with back propagation (BP). In both accuracy and computational time, BP performed better than AIS. However, the author claimed that the AIS algorithm is comparable with BP.

Hsu et al. used a fuzzy expert system to forecast the hourly load of the Taiwan Power Company [94]. Ranaweera et al. used the data from a large metropolitan utility in North America to demonstrate that the load forecasting accuracy of fuzzy logic models was comparable to more complicated statistical and ANN methods [95]. Similar experiments have been done by Pandian et al. with electricity load data from the Neyveli Thermal Power Station in India [96].

Ahmadi et al. found fuzzy logic more accurate and faster than conventional methods for short-term load forecasting [97]. However, only two 24-hour forecasts were shown in this article, and no comparisons with conventional methods were presented. Sunandaj Power Network in Kurdistan of Iran was considered as a case study for this experiment.

Gradient boosting is a machine learning technique for regression and classification [98], which can be used as a prediction model. A set of weak prediction models are combined to create a strong prediction model. This technique works well due to its strong resistance to overfitting [99]. Taieb et al. [51] used a gradient boosting approach for solving a hierarchical load forecasting problem in GEFCom2012 [47]. The forecasting problem had 20 zones and 11 weather stations. Non-parametric additive models [100] were used for forecasting each zonal load. Temperature effect, calendar effect, and lagged demand effect were considered in the forecasting model. Data were analyzed for outlier detection and cleaning. 24 hourly models were generated for forecasting each hour of the day. This technique ranked fifth among 105 participating teams in GEFCom2012. Similar kinds of boosting techniques were applied by Hyndman et al. [101] to forecast short term and long term electricity demand for the Australian Electric Power System.

2.2 Probabilistic Forecasts

This section provides background knowledge of probabilistic forecasting methods, which are used to quantify forecast uncertainties in Section 3.2. Probabilistic forecasting is a century old problem. The first article on probabilistic forecast was published in 1906 by Cooke [102]. It considered uncertainty inherent in weather forecasts. A comprehensive discussion of probabilistic forecasting between 1900 and 1980 considering purpose and use of probability forecasts, procedures, probability forecast types, evaluation techniques, and problems was prepared by Hughes (1980) for

the U.S. National Weather Service (NWS) [103]. Precipitation was used as the case study. Murphy et al. (1984) studied the use of probabilistic forecasts in meteorology between 1900 and 1984 [40]. In 1965, precipitation probability was added to all public weather forecasts, which was the first practical use of probabilistic forecasts in any field. Since then, probabilistic forecasts have been used in other fields such as economics, medicine, decision analysis, and risk analysis (earthquake and nuclear power plant) [40]. However, probabilistic forecasts have only recently been introduced in the energy sector [37].

Subsection 2.2.1 presents reviews of probabilistic forecasts. Recently published probabilistic forecasting methods are reviewed in two subsequent subsections: statistical approaches (see subsection 2.2.2) and machine learning approaches (see subsection 2.2.3).

2.2.1 Probabilistic Forecasting Review

Chatfield (1993) reviewed the importance of interval forecasting, which is a special case of probabilistic forecasting [104]. In the tutorial section, the author discussed several general methods to calculate predictive intervals (PI) such as 1) evaluating the variance of forecasting error based on an assumption of normality, 2) fitting a probability model, 3) using ARIMA and exponential smoothing methods assuming optimality, 4) using “approximate” formulas, 5) calculating empirical error distributions, 6) simulating and resampling error distribution, 7) Bayesian approach, and 8) calculating PI from transformed variables. The variable transformation technique is applied in one of the

methods in this dissertation to generate probabilistic forecasts (see Section 3.4). Chatfield updated his review in 2001 [105]. Chatfield published a third review in 2013, in which he compared time series forecasting methods, including point and interval forecasts [106]. Most of the review was related to recent developments in forecasting methods.

Tay et al. (2000) surveyed the use of density forecasts (also known as probabilistic forecasts) in macroeconomics and finance. Density forecasts of U.S. and UK inflation rates were presented as examples. Tay et al. stressed the importance of the presentation of density forecasts. A fan chart was shown as an illustration. Presentation of the probabilistic forecasts is important to avoid misunderstanding. Assessment techniques of density forecasts were discussed briefly. Probabilistic Integral Transform (PIT) is mentioned as a graphical tool to measure calibration of the probabilistic forecasts. Raftery (2014) took a similar approach to demonstrate the use of probabilistic forecasts in different fields of research [107]. This article discussed five successful practical uses of probabilistic forecasts. Although the communication of probabilistic forecasts is more difficult than the communication of point forecasts, a graphical approach was preferred to a tabular format to present probabilistic forecasts in most of the cases. Spiegelhalter et al. reported on new ways to visualize forecast uncertainty [108]. In this work, the assessment of the probabilistic forecasts is presented in both numerical and graphical ways.

Zhang et al. (2012) reviewed state-of-the-art methods and recent developments in wind power probabilistic forecasts [109]. Wind power uncertainty forecasting was classified based on time scale and application: very short-term (seconds to minutes), short-term (hours to days), medium-term (days to weeks), and long-term (weeks to

months to years). The impact of wind power forecasts on electricity prices was demonstrated qualitatively in the supply-demand curve. The uncertainty of wind power can be expressed in four ways: a) PDFs and CDFs, b) quantiles and intervals, c) discrete probabilities, and d) moments of probability distribution (mean, variance and skewness). The authors classified wind power forecasting based on mathematical methods: parametric (homoscedastic time-series, heteroscedastic time-series, and artificial intelligence), and non-parametric (quantile regression, kernel density estimation, ensemble forecasting, and artificial intelligence). In the field of parametric approaches, most research has been done in four areas: a) shape assumption of predictive distributions, b) estimator of location parameters, c) estimator of scale parameters, and d) parameter evaluation theory. Different variants of quantile regression models and kernel density estimators were discussed in that paper. Three criteria were mentioned as required properties of probabilistic forecasts: a) reliability, b) sharpness, and c) skill score. Reliability is defined by the statistical consistency of a predictive distribution. Calibration replaces reliability as an important assessing criteria of probabilistic forecasts in some articles [110, 111]. Calibration is defined as the statistical compatibility between probabilistic forecast and observations. Figure 2.1 provides examples of perfectly reliable and well calibrated probabilistic forecast (left), and non-reliable and poorly calibrated probabilistic forecast (right), respectively. In this work, a new scoring rule is used, where reliability is considered as the most important criteria to assess probabilistic forecasts.

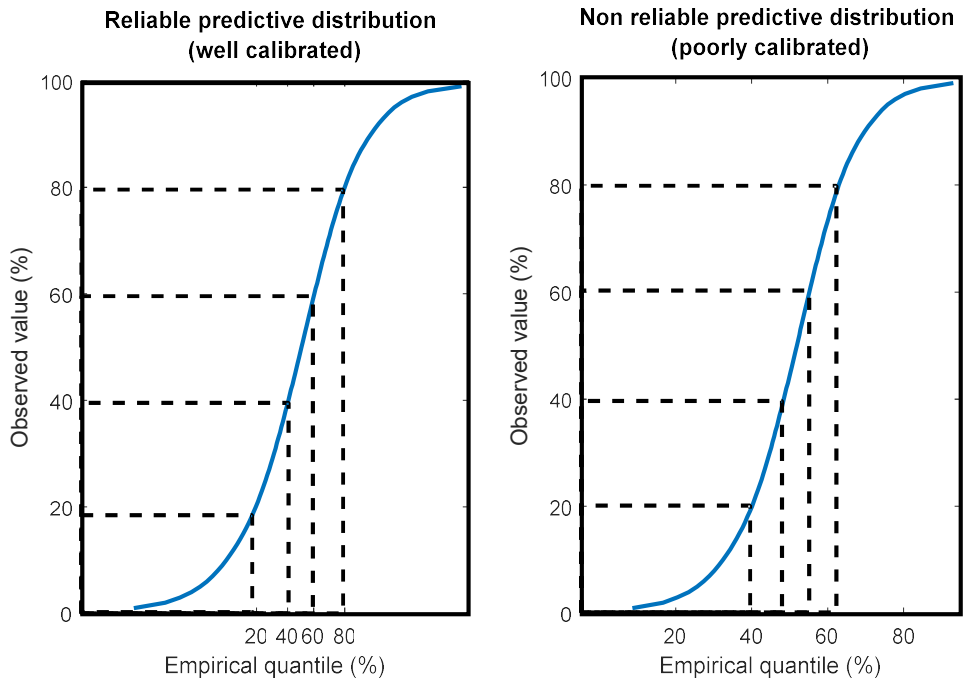

Figure 2.1: Reliability of probabilistic forecasts.

Sharpness denotes the concentration of a predictive distribution. Sharper probabilistic forecasts are considered useful and preferable in the GEFCom2014 [112]. However, too sharp probabilistic forecasts are not desirable. Figure 2.2 illustrates an instance of sharper and less sharp probabilistic forecasts compared with a reference probabilistic forecast.

Skill scores referred to different scoring rules to assess a predictive distribution. A proper scoring rule provides the best score by forecasting the true distribution [113]. Thus a proper scoring rule is preferable. A list of widely used skill scores such as logarithmic score, continuous ranked probability score (CRPS), trick or check loss score, interval

score (IS) (also known as Winkler score [114]), and energy score were discussed in that article. CRPS is used in this work to assess probabilistic forecasts.

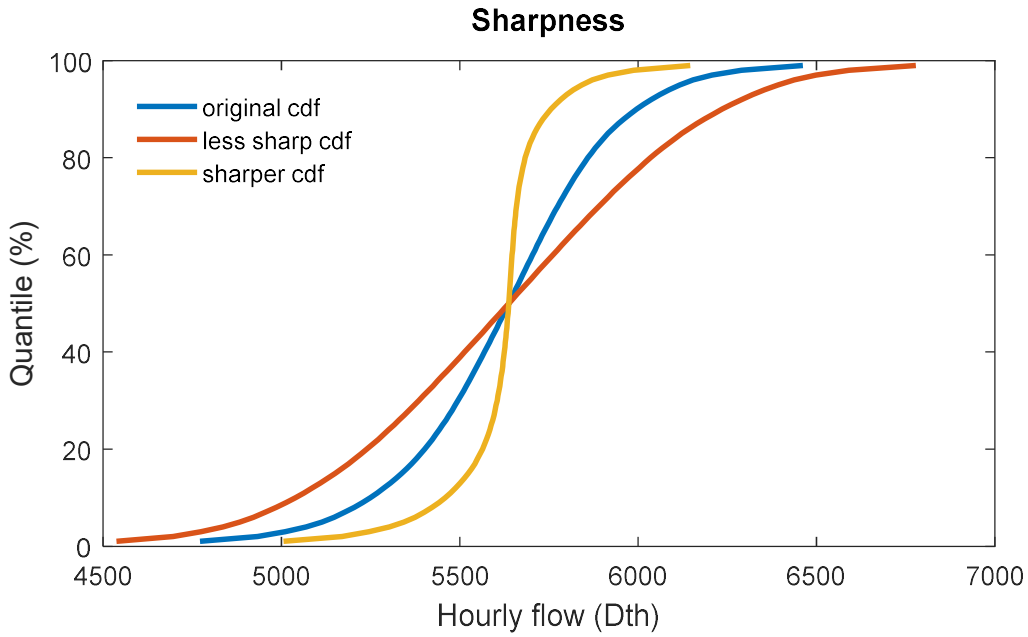

Figure 2.2: Sharpness of probabilistic forecasts.

Pinson (2013) reviewed probabilistic wind power forecasting as part of operational management for the electricity market [115]. Wind power generation was shown to be a stochastic process. The author argued the importance of using probabilistic forecasts over point forecasts in the energy industry. A Western Denmark dataset was used as a case study to illustrate sample probabilistic forecasts and space-time trajectories. Probabilistic forecasts were presented in a colorful way similar to the bank of England's fan chart [116]. The author related the operational decision-making problem of the power industry with a variant of the well-known linear terminal loss problem, also known as the newsvendor problem [117], where probabilistic distributions of demand are

required to make optimal decisions [118, 119, 120]. Four future challenges for probabilistic forecasts were discussed at the end of this article: (a) improving forecasts by extracting more out of the data, (b) new forecasting methods, (c) evaluating probabilistic forecasts, and (d) bridging the gap between forecast quality and value. This dissertation tries to meet all four challenges mentioned in this literature.

Weron (2014) reviewed a variety of models (from 801 papers) used in electricity price forecasting (EPF) between 1989 and 2013 [121]. The article focused on describing the existing solutions of EPF, with special emphasis on the strengths and weaknesses of the individual models. The author categorized all reviewed articles into five groups based on forecasting methods to find the most popular one: a) multi-agent models (Nash-Cournot framework, supply function equilibrium, strategic production-cost, agent-based), b) fundamental and structural models (parameter rich fundamental, parsimonious structural), c) quantitative and stochastic models (jump-diffusions, Markov regime-switching), d) statistical approaches (similar day, exponential smoothing, regression analysis, AR/ARX- type, GARCH-type), and e) machine learning techniques (feed-forward neural networks, recurrent neural networks, fuzzy neural networks, SVM). Historical electricity price datasets were collected from the Nord Pool power market (Denmark, Finland, Norway, and Sweden) [122] to illustrate the dynamic changes in electricity price over time. Interval forecasts, probabilistic forecasts, and combining forecasts were mentioned as the future of EPF beyond point forecasts. Probabilistic forecasts were hardly used in EPF. Probabilistic forecast evaluation was mentioned as more complicated than point forecast evaluation. A guideline for evaluating forecasts

were given at the end of this paper. The new evaluation technique, graphical calibration score (GCM), implemented here in this work is motivated by useful guidelines and two evaluation criteria (resolution and sharpness) mentioned in this paper.

Hong et al. (2014) [37] reviewed probabilistic electric load forecasting from two perspectives: applications and methods. The authors studied probabilistic forecasting as an application for a) generating probabilistic load forecasts considering uncertainties of generation outages, changes in network configuration, and load forecasting error, b) minimization of unit production cost considering production uncertainty and other constraints on the units and the system, and c) reliability planning. Different methods of probabilistic forecasting were reviewed in three separate sections: short-term (hourly/daily), long-term (monthly/annual), and interval forecasting without probabilistic meaning. The article also covered an extensive literature review of short-term point load forecasting based on techniques, methods, and significance. In the tutorial section, the authors mentioned three different ways to generate probabilistic forecasts with useful infographics: a) input data simulation to generate different scenarios, b) model-driven probabilistic forecasts, and c) error simulation from point forecasts and ensembles of point forecasts. The authors mentioned a lack of well-established evaluation tools as one of the main reasons for the under-development of probabilistic forecasts in the energy sector. Reliability, sharpness, and resolution were mentioned as the main criteria to evaluate probabilistic forecasts. In this work here, a new probabilistic forecast evaluation technique is implemented considering above mentioned criteria. The Pinball loss function and the Winkler score were discussed as comprehensive metrics for probabilistic

forecasts. In this dissertation, the Pinball score is used to assess probabilistic forecasts. Eight traditional problems with probabilistic forecasts were identified in that article: a) reproducible and comprehensive studies, b) leveraging the point forecasting articles, c) scenario generation for probabilistic forecasts, d) error measurement for probabilistic forecasts, e) probabilistic forecast combination methods, f) hierarchical probabilistic forecasts, g) high performance computing, and h) valuation of probabilistic forecasts improvement. Five new problems in the energy industry that have introduced a great deal of uncertainty were mentioned in that article: a) climate variability due to global warming, b) growing market of electric vehicles, c) wind and solar power generation, d) replacing old electric equipment with new energy efficient equipment, and e) dynamic electricity price.

Gneiting et al. (2014) offered an overview of the probabilistic research field by providing theory, methods, evaluation metrics, and applications [111]. The importance of using probabilistic forecasts in different fields of research were explained using illustrative examples such as the Bank of England's probabilistic inflation rate forecasts and Stateline wind energy center's probabilistic wind speed forecasts. Sharpness and calibration were mentioned as the criteria for good probabilistic forecasting. The probabilistic integral transform (PIT) was shown as a tool to measure the calibration of probabilistic forecasts. A scoring rule should measure both sharpness and calibration. Various scoring rules to measure the quality of probabilistic forecasting including quadratic score (QS), logarithmic score (LS), Hyvärinen score (HS), David-Sebastiani score (DSS), continuous ranked probability score (CRPS), and interval score (IS) were

explained and compared in a case study of the Stateline wind speed forecasting. The CRPS was mentioned as a proper scoring rule. CRPS is used as one of the scoring rules to assess probabilistic forecasts for this dissertation. The importance of ensemble forecasts in different applications has been increased recently. Two state-of-the-art statistical methods, nonhomogeneous regression (NR) and Bayesian model averaging (BMA), were applied to generate 24-hour ahead probabilistic weather (temperature, wind speed, and precipitation) forecasts for Frankfurt, Germany. No evaluation technique was applied to assess the probabilistic forecasts.

Rossi (2015) discussed the usefulness of employing density forecasts (also known as probabilistic forecasts) in economics forecasting and policymaking [123]. Recent development in density forecasts were reviewed. A rolling estimation scheme for producing forecasts was presented. A density forecast of the U.S. real gross domestic product (GDP) growth during the financial crisis (2008) was demonstrated using various visual presentations, including a fan chart. Available evaluation techniques for measuring the quality of density forecasts were studied in that article. The performance of a U.S. real GDP growth density forecast was measured using probability integral transforms (PIT). PIT was introduced by Diebold et al. in 1998 [124].

In 2014, the IEEE Power and Energy Society sponsored a global energy forecasting competition (GEFCom2014) on the probabilistic forecasting theme to identify the best probabilistic forecasting methods in four tracks: load forecasting, price forecasting, wind forecasting, and solar forecasting. Hong et al. (2016) [112] gave an overview of the GEFCom2014 and summarized the top five winning strategies from each

track. The pinball loss function was used as a scoring rule to measure the performance of all forecasts submitted for this competition. The pinball loss function is used in this dissertation to assess proposed probabilistic forecasting methods. In both electricity and price forecasting tracks, the winning team used generalized additive models with quantile regression [125]. The most useful part of that paper is the table containing different techniques used by different teams for each track. All datasets used for this competition were made publicly available.

Similar to point forecasting methods, probabilistic forecasting methods are discussed in two sections. The next subsections review statistical and machine learning methods for generating probabilistic forecasts.

2.2.2 Statistical Methods for Producing Probabilistic Forecasts

Although it is possible to trace back probabilistic forecasting articles more than one hundred years [102], the use of probabilistic forecasts in the energy sector is new compared other areas of research such as meteorology, economics, and finance. Most of the development of probabilistic forecasts in the energy sector has been done within the last five years [37]. In 1992, the Electric Power Research Institute (EPRI) arranged an energy forecasting competition to develop a better one day-ahead hourly energy forecasting model than the existing model for Puget Sound Power and Light Company (PSE) [126]. Eleven participating teams were asked to generate probabilistic forecasts based on eight years (1983-1990) of historical hourly electricity usage and actual temperatures provided by PSE. This contest is the first documented use of probabilistic

forecasts in the energy industry. The Quantitative Economic Research Inc. (QUERI) prepared a report in 1993 on behalf of EPRI for PSE. Their report documented the forecasting competition and the selected probabilistic method for forecasting hourly loads for PSE [127]. The winning method used an ordinary least squared (OLS) method to analyze historical errors. Through investigation, QUERI found that using hourly models provide improved forecasts compared to a single model for every hour. Thus, 48 hourly models (24 for weekdays and 24 for weekends) were used. Probabilistic forecasts were indicated as a better option than point forecasts for decision making in the energy industry. The QUERI report has special importance to understand how forecasts are used in the U.S. electricity industry for daily decision making because of the direct involvement of the oldest U.S. utility, PSE (established in 1873 [126]).

Hong et al. proposed a scenario-based long-term probabilistic load forecasting (LTLF) model [128]. Based on 30 years of hourly weather information and three forecast macroeconomic scenarios (base, aggressive, and conservative) for one year, they have generated 90 cross scenarios. The authors used the electric load of North Carolina Electric Membership Corporation (NCEMC) as a case study to generate one year ahead probabilistic load forecasts based on different scenarios. Four MLR models, similar to the MLR benchmark model of Hong's Ph.D. dissertation [45], were used to generate temperature scenario based forecasts (implemented in SAS [129]). The performance of the probabilistic forecasts was not evaluated.

Xie et al. simulated forecast residuals to generate long term probabilistic load forecasts based on an assumption of normality [130]. The Kolmogorov–Smirnov

normality test was used to check normality. Weather scenarios were generated from historical weather data [128]. Three linear models were used to generate point forecasts based on different weather scenarios. Then probabilistic forecasts were calculated from a set of point forecasts. Two case studies (NCEMC and the GEFCom2014 dataset [112]) were used to produce probabilistic forecasts. The performance of the probabilistic forecasts was evaluated by the pinball loss function. The normality assumption should be avoided in practice because the error distribution typically is not normal. Similar techniques as in [128, 130] (temperature scenario generation from historical weather data) were adopted by Xie et al. to produce probabilistic forecasts in the GEFCom2014 electricity forecasting track [131]. Data were pre-processed in two steps: a) weather station selection [54] and b) data cleaning.

Pierrot et al. (2011) applied a semi-parametric method using Generalized Additive Models (GAM) to forecast short-term electricity load for a French electricity company, Electricité De France (EDF) [132]. GAMs are regression models that use smoothing splines instead of linear coefficients. GAMs are very effective in capturing non-linear effects [133]. The electricity demand reduced significantly during the summer break in EDF, which was well captured by the GAM. The model was fitted such that the Generalized Cross Validation (GCV) score (lower score is better, included in the GAM package of R [134]) is minimized, leading to nearly normal forecasting errors. The overall performance of this method was competitive with the existing EDF operational model. However, utilities are often more concerned about special days rather than overall

performance of a model. Thus, this dissertation calculates the performance of probabilistic forecasting methods on usual days [135].

Wood et al. applied the same model (GAM) on larger dataset (five years of half-hourly electricity demand) collected from the same French utility, EDF, to forecast electricity demand [136]. The main goal of this experiment was to show GAM working for a larger dataset. A similar approach (using GAM) was used by Fan and Hyndman to forecast half-hourly load for seven days [137] and long-term peak electricity [138] for South Australia, respectively. Australian Energy Market Operators (AEMO) have used a short-term load forecasting model to forecast half-hourly load for Victoria and South Australia. For the long term density forecasts, 2000 years of temperature scenarios were generated from simulation to generate probability distributions of weekly and yearly peak load. The authors assert good performance of the probabilistic forecasts without comparing or scoring the forecast using any proper scoring rule. Dordonnat et al. used a semi-parametric regression model similar to [132] in the GEFCom2014 probabilistic load forecasting track [139]. They have selected three weather stations among given twenty-five based on the lowest GCV score. Temperature scenarios were generated by simulating different paths between normal temperature (moving average) and an AR model. Probabilistic forecasts were produced from 1000 simulated temperature scenarios using a GAM model. Forecast performance was evaluated using the pinball loss function according to the competition rules. The proposed method ranked among the top five methods in the load forecasting track (as team ADADA) of GEFCom2014 [112].

Combinations of point forecasts provide better forecasts than individual point forecasts [140, 141, 142, 143, 144]. However, ensembles of probabilistic forecasts have not been explored [37]. Quantile regression (QR) [145, 146, 147, 148, 149, 150], which measures the conditional distribution of the dependent variable given explanatory variables, was used recently in the energy sector to generate probabilistic forecasts from several point forecasts. Nowotarski et al. examined the idea of combining forecasts to generate a day-ahead electricity spot price probabilistic forecast using quantile regression averaging (QRA) [151]. The Jersey Central Power and Light Company (JCPL) of the Pennsylvania-New Jersy-Maryland (PJM) interconnection was used as a case study for this paper. The QRA method combined twelve individual point electricity price forecasts to generate 50% and 90% interval forecasts. The authors have shown better forecasting performance of the QRA method by comparing with individual points forecasts. Liu et al. proposed a similar idea to generate probabilistic electricity load forecasts by combining a set of sister models through quantile regression averaging (QRA) [152]. Publicly available data from GEFCom2014 was used to show the effectiveness of this method. The main dataset was divided into eight non-equal parts to create eight different training datasets, and the forecasting models generated from those individual datasets were called sister models. The pinball loss function [153] and Winkler scores [114] were used to assess the probabilistic forecasts. The authors compared the combined probabilistic forecast result with individual model forecasts, where the QRA outperformed all individual model forecasts.

Team Poland, which finished second in the GEFCom2014 probabilistic electricity spot price forecasting track [112], applied a similar strategy of combining point forecasts using QRA [151]. Twenty-four separate models were constructed to capture the pattern of different hours of the day. Maciejowska et al. explained Team Poland's electricity price forecasting strategy in [154]. The same authors extended the QRA method by using principal component analysis (PCA) to select point forecasting models automatically from a set of 32 individual forecasting models, which was called Factor Quantile Regression Averaging (FQRA) [155]. Data collected from the British power market were used to generate 24-hour ahead probabilistic electricity load and price forecasts to show the effectiveness of this method. The forecasting performance of FQRA, QRA, and an individual AR-type model (used as a benchmark) were compared using a Winkler score. The FQRA outperformed other two methods most of the time.

Gaillard et al. proposed a method based on GAM and quantile regression (QR), called quantGAM, to forecast electricity load and price, which ranked first (as team Tololo) in both the load and the price forecasting tracks of GEFCom2014 [156]. GCV score was used to select the top four weather stations from twenty-five weather stations. 800 randomly generated temperature scenarios were fed into GAM to generate scenario-based point load forecasts. Point forecasts were used as inputs to QR to generate probabilistic load forecasts. In the electricity price forecasting track, two other probabilistic forecasting methods were proposed and compared with quantGAM. The first method was based on the idea of combining individual predictors such as AR-type models, linear regressions, GAM, random forests regressions, and gradient boosting

machines using QRA, similar to [151]. The second method was kernel-based QR with lasso penalty for covariate selection, denoted as quantGLM. Haben et al. used kernel density estimation (KDE) and QR in the load forecasting track of GEFCom2014 [157], similar to the third model mentioned by team Totolo in the price forecasting track. They did not consider any kind of preprocessing approach like other winning teams, such as data cleaning, weather station optimization, and temperature scenario generation.

Jeon et al. proposed two statistical methods, ARMA-GARCH and conditional kernel density estimation (KDE) to produce 24-hour ahead probabilistic wave energy forecasts [158]. Three types of data transformation methods, log, square root, and Box-Cox transformations, were used to convert data into a normal distribution and compared with a no data transformation technique using CRPS. Log, and Box-Cox transformation methods produce the best score. Datasets for this experiment were collected from the FINO1 research platform located in the North Sea near Germany [159]. An illustrative comparison between regression methods, AMRA-GARCH models, and conditional kernel density estimation (KDE) have been shown. PIT histograms were used to demonstrate the calibration of the probabilistic forecasts.

Mangalova et al. used a nonparametric approach based on the Nadaraya-Watson estimator for short term probabilistic load forecasting, which ranked fifth in GEFCom2014 [160]. This method does not require any assumption about the probability distribution. The authors further modified the transformation process of quantiles after the competition, which led to better probabilistic forecasts than those submitted to GEFCom2014.

Ziel et al. presented a method based on lasso (least absolute shrinkage and selection operator) estimator [161] for short term probabilistic load forecasting [162]. The proposed method was applied on two publicly available datasets (GEFCom2014-L and GEFCom2014-E) [112] and compared with two benchmarks [49, 50]. The pinball score was used for evaluating probabilistic forecasts. Lasso estimator performed better than four benchmark models based on pinball score. The pinball score is used in this dissertation to evaluate probabilistic forecasts.

Most of the probabilistic forecasting methods are statistical. However, there are a few articles that use machine learning approaches with statistical methods. The next subsection presents machine learning approaches for generating probabilistic forecasts such as gradient boosting machines, ANN, *k*-NN clustering, radial basis function (RBF), support vector machine (SVM), particle swarm optimization (PSO), and decision trees.

2.2.3 Machine Learning Methods for Producing Probabilistic Forecasts

Taieb et al. proposed 24-hour ahead 50% and 90% load prediction intervals from high frequency smart meter data using a boosting technique with additive quantile regression [163]. Regression trees, smoothing splines, and penalized regression splines (P-spline) were used as base learners in the boosting algorithm. This method was compared to three benchmark models for both aggregated and disaggregated scales. The continuous ranked probability score (CRPS) [113] was used as a scoring rule for the comparison.

Similar machine learning approaches such as gradient boosting machines (GBM) with a quantile loss function have been used by Landry et al. for short term probabilistic wind power forecasting [164]. Standard smoothing techniques and a cross-sectional approach were applied to adapt with forecast inaccuracies. The pinball loss function was used to measure the performance of their probabilistic forecasts. Their proposed method secured the top position in the GEFCom2014 wind forecasting track. However, the distribution of probabilistic forecasts was concentrated near the 40^{th} quantile (more than 60% of observations were between the 20^{th} and 50^{th} quantiles, and less than 10% of observations were between the 60^{th} and 99^{th} quantiles), which indicates that the proposed method did not capture the true distribution. Forecasts are more effective if they can perform reasonably during extreme conditions (tail of the distribution). Less than 1% of observations between the 70^{th} and 99^{th} quantiles proved the weakness of this method to produce credible forecasts during extreme conditions. Still, the probabilistic forecast generated by this method scored well using the pinball loss function, which raises questions about the credibility of the scoring rule used in GEFCom2014. This dissertation proposes two new scoring rules, which assign a better score if the empirical distribution matches with expected distribution and penalize heavily for being extra sharp or less sharp.

Nagy et al. investigated short-term probabilistic solar and wind power forecasting using two machine learning techniques, voted ensemble of quantile regression forecasts (QRF) and stacked random forecasts – gradient boosting decision tree (GBDT) [165]. The probabilistic forecasts obtained from the above methods were post-processed by

isotonic regression to maintain the monotonic-increase attribution of probability distributions. This approach ranked second in both wind and solar forecasting tracks of GEFCom2014 [112].

Juban et al. used a multiple quantile regression approach to predict probabilistic wind, solar, and electricity price in GEFCom2014 [166]. The proposed method used a radial basis function (RBF) to capture the non-linear dependencies on the input data, and a *k*-means clustering algorithm was used to compute the center of a RBF. The alternative direction method of multiplication (ADMM) was used for solving the optimization problem resulting from multiple quantile regression. This method performed in the top five of GEFCom2014 wind, solar, and electricity price forecasting tracks.

Quan et al. developed a particle swarm optimization (PSO) based ANN model for quantifying uncertainties associated with load forecasts, called LUBE (lower upper bound estimate) [167]. Historical load data collected from Singapore, Ottawa (Canada), and Texas (USA) were used as case studies to demonstrate 168 hours (1 week) ahead probabilistic load forecasts. The proposed method was compared with three benchmark models (ARIMA, ES, and naïve model, which is similar to the point forecasts). Four evaluation metrics including three predictive interval (PI) width assessment indices and the Winkler score [114] were used to measure the performance of probabilistic forecasts. Their proposed method is compared to three benchmarks. The ANN also used by Dudek for short term probabilistic forecasting ranked third in the GEFCom2014 wind power forecasting track [168]. The author has shown a high correlation between electricity price and recent load demand. Thus, the proposed method only considered recent load demand.

One of the weaknesses of this method is the normality assumption of the error distribution, which is one of the easiest ways to generate probabilistic forecasts but is not practical. In that work, a benchmark model is created based on normality assumption to demonstrate why this approach should be avoided.

Zhang et al. used *k*-Nearest Neighbor (*k*-NN) and a kernel density estimator (KDE) for short-term solar power forecasting [169]. First, a *k*-NN algorithm was used to group days with similar weather conditions. Then, the KDE was applied to produce probabilistic forecasts for each of the groups created by *k*-NN. The authors used this method on the GEFCom2014-S dataset [112] to demonstrate its effectiveness. The performance of this method has been tested using the pinball loss score. However, the result was not compared with any benchmark method. Similar approaches to find the probabilistic distribution ranked fifth in the GEFCom2014 wind power forecasting track [170]. A *k*-NN algorithm was used also by another two teams in GEFCom2014. In the probabilistic wind power forecasting track, Mangalova et al. defined wind speed as a distance metric for the *k*-NN algorithm, and then linearly interpolated each quantile to calculate quantile estimation [171]. That method ranked third in the wind power forecasting track.

Huang et al. applied gradient boosting to find point forecasts, and then *k*-NN regression was used to generate probabilistic forecasts by finding similar scenarios from the historical data [172]. This method ranked first in the solar power forecasting track.

2.3 Probabilistic Forecast Evaluation Techniques

One of the major contributions of this dissertation is providing a new way to assess probabilistic forecasts. Some of the existing evaluation techniques are used in the dissertation to assess probabilistic forecasting methods described in Section 3.2. Thus, this section focuses on providing enough background information of existing evaluation techniques from different fields of research.

Measuring forecast performance is important to identify weak points of a model, which leads to improved forecasting models. Different types of evaluation techniques used to assess point forecasts were reviewed by Hyndman et al. in [173]. Among all available evaluation techniques, mean absolute percentage error (MAPE) and mean absolute error (MAE) are widely used by the research community for assessing the quality of point forecasts. However, there is no unique popular evaluation technique available for evaluating probabilistic forecasts, which is considered as one of the main reasons for the slow progress of probabilistic forecasting research [37]. A good probabilistic forecast aims to maximize the sharpness of its predictive distribution, subject to calibration [174]. Reliability and resolution are important criteria [37, 175]. Gneiting et al. reviewed some of the well-known evaluation techniques to evaluate probabilistic forecasts in [111]: quadratic score (QS), logarithmic score (LS), Hyvärinen score (HS), David-Sebastiani score (DSS), continuous ranked probability score (CRPS), and interval score (IS). The rest of the chapter provides an overview of some well-known probabilistic forecast evaluation techniques.

In 1950, Brier proposed an evaluation technique to evaluate probabilistic rain forecasts [176]. The Brier score or probability score (PS) is considered the earliest attempt to score a probabilistic forecast [103].

Murphy presented a vector partition for the PS, which consists of three terms: a) a measure of uncertainty, b) a measure of reliability, and c) a measure of resolution [177]. Several versions of PS were studied by Murphy, including skill score (SS) [178], collective skill score (CSS) [179], and sample skill score (SSS) [180]. Hernandez et al. presented a graphical version of PS, called the Brier curve or the receiver operating characteristic (ROC) curve [181, 182]. PS is considered as one of the best scoring rules to evaluate probabilistic forecasts by most meteorologists [183].

A weighted version of PS, called ranked probability score (RPS), performs better than PS for evaluating distance-sensitive probability distributions such as temperature forecasts [183]. However, PS or RPS is not suitable for evaluating continuous probability distributions. Gneiting et al. proposed a continuous version of PS/RPS, continuous ranked probability score (CRPS). CRPS is credible to measure calibration and sharpness of a probability distribution [110]. The CRPS has been shown to be a proper scoring rule in [113]. If n is the number of forecasts, $F_i^f(x)$ denotes the forecasted cumulative distribution function (CDF), and $F_i^0(x)$ represents observed value, then

$$CRPS = \frac{1}{n}\sum_{i=1}^{n}\int_{x=-\infty}^{x=\infty}\left(F_i^f(x) - F_i^o(x)\right)^2 dx \ . \tag{2.1}$$

A lower CRPS score is considered better. This evaluation technique has been used for evaluating probabilistic forecasts in [107, 111, 158, 163]. This technique yields a better score for sharper but low calibrated probability distributions compared with less sharp but well calibrated probability distributions. However, calibration should be given more importance than sharpness for evaluating probabilistic forecasts, because it penalizes too sharp probabilistic forecasts. Two scoring rules, quantile calibration score (QCS), and percentage quantile calibration score (PQCS), providing calibration as the highest priority are defined in Section 3.5.

Diebold et al. proposed an evaluation technique to measure calibration of probability distribution called the probability integral transform (PIT), which is well accepted by most of the economics literature [124]. This technique provides a visual representation of calibration. If the PIT looks like a uniform distribution, the density forecast is well calibrated. This evaluation technique has been adopted (especially in the financial sector) as a probabilistic forecast measuring tool [111, 123, 124, 158, 163, 110, 184, 185]. One problem with this evaluation technique is not providing any numerical score, which is not helpful to compare two very closely related probability distributions. The new evaluation technique, graphical calibration measure (GCM), presented in Section 3.5 has both graphical and numerical representations.

Different kinds of loss functions, which can be used to evaluate probabilistic forecasts including linlin, hinge, tick, pinball, and newsvendor loss, were studied by Gneiting in [120]. Recently, the pinball loss function [153] and the Winkler score [114] have been used by several articles to evaluate probabilistic forecasts [130, 152, 163, 167].

Pinball and Winkler scores can measure reliability and sharpness of probabilistic forecasts [37]. Both of them are proper scoring rules. The pinball loss function has been selected as an official scoring rule in a major global forecasting competition, GEFCom2014 because of its simplicity [112].

If q represents forecast quantiles (for example: 0.01, 0.02, … , 0.99), y_t is the actual flow at time t, and $\hat{y}_{t,q}$ denotes forecasted flow for quantile q at time t, the pinball score for an individual quantile is

$$Pinball = \begin{cases} (1-q)\left(\hat{y}_{t,q} - y_t\right), & y_t < \hat{y}_{t,q} \\ q\left(y_t - \hat{y}_{t,q}\right), & y_t \geq \hat{y}_{t,q} \end{cases}. \tag{2.2}$$

If L_t and U_t represent lower and upper bounds of a predictive interval at time t and δ is the difference between U_t and L_t , the Winkler score is

$$Winkler = \begin{cases} \delta, & L \leq y_t \leq U \\ \delta + 2\left(L_t - y_t\right), & y_t < L_t \\ \delta + 2\left(y_t - U_t\right), & y_t > U_t \end{cases}. \tag{2.3}$$

For both pinball and Winkler scores, lower scores are better. Both scoring rules give more importance to sharpness than calibration (sharp forecast with low calibration always achieve better scores than less sharp with well calibrated forecast), which should be opposite in practice [164].

Other evaluation techniques such as the Kolmogorov-Smirnov (KS), the Cramér–von Mises test statistic [186], and Monte Carlo simulation [187] have been applied to assess probabilistic forecasts. However, use of these tests are limited.

This chapter presented a literature review on two different ways of forecasting: point forecasting and probabilistic forecasting. Highly cited literature surveys were included at the beginning of this chapter to cover most of the historically important forecasting methods concisely. Recent articles were discussed in two sections: statistical approaches and machine learning approaches. A short review of existing evaluation techniques for assessing forecasts was included at the end of this chapter. The next chapter of this dissertation presents a point forecasting method, three probabilistic forecasting methods (including a benchmark), and an evaluation technique to assess probabilistic forecasts.

CHAPTER 3

PROBABILISTIC FORECASTING METHODS AND EVALUATION TECHNIQUES

This chapter presents the two major contributions of this dissertation: a fast and efficient way to quantify forecast uncertainty using the Johnson curve [188] and a new evaluation technique to assess probabilistic forecasts. The point forecasting method used for generating probabilistic forecasts also is presented. A benchmark probabilistic forecasting model was created based on a normality assumption, as explained in Section 3.2. A newly developed unpublished probabilistic forecasting engine using a kernel density estimator (KDE) is introduced in this chapter [189]. Three variants of the Johnson curve and KDE based probabilistic forecasting methods are presented in Sections 3.3 and 3.4, respectively. This chapter ends with a new evaluation technique, graphical calibration measure (GCM) to assess probabilistic forecasts. Associated with GCM are two new scoring rules, quantile calibration score (QCS) and percentage quantile calibration score (PQCS).

3.1 Point Forecast Using Multiple Linear Regression

This dissertation generates probabilistic forecasts through error analysis of point forecasts. A linear regression is used to generate point forecasts from weather inputs, seasonal effects, and historical load demand. This section describes the point forecasting method used to generate the probabilistic forecasts. However, the probabilistic

forecasting methods presented in this dissertation will work with any point forecasting method, this one is only intended as an example.

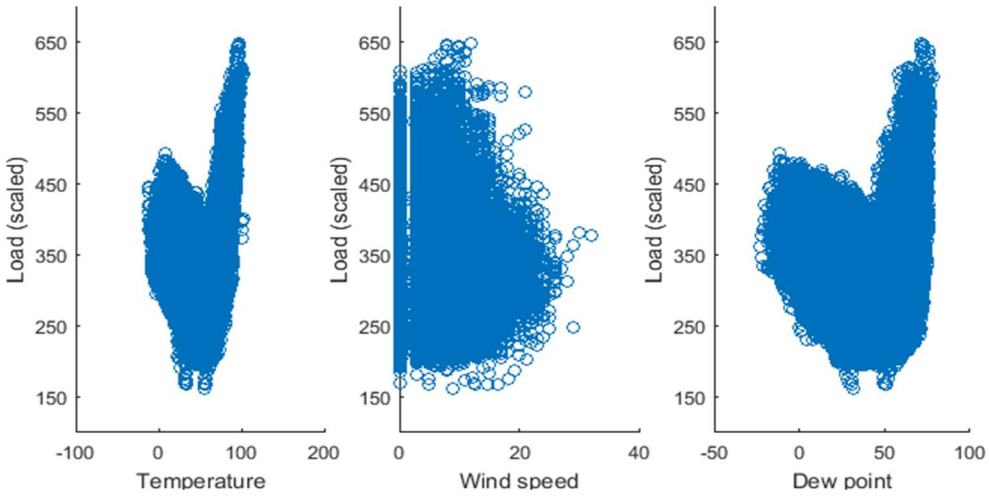

Figure 3.1: Electricity load vs. weather inputs.

The linear regression model used in this dissertation is based on Siddique's work, MLR1 [46], and Deoras' work, MLR2 [78]. The final linear regression model considers an hourly holiday effect (8 A.M. to 4 P.M.) and lag terms (1-3, 23-25, and 167-169 hours). Energy demand is highly related to weather inputs such as temperature, wind speed, and dew point (see Figure 3.1). Therefore, these three weather inputs are used in the implemented multiple linear regression point forecasting method, MLR3. Sine and cosine terms of the day of the week and the day of the year are considered as input factors of the MLR3 model to capture weekly (see Figure 3.2) and yearly cycles (seasonality), respectively. Table 3.1 compares the factors of the MLR3 method with the initial two LR methods, MLR1 [46] and MLR2 [78]. In principle, any other point forecasting method could be used.

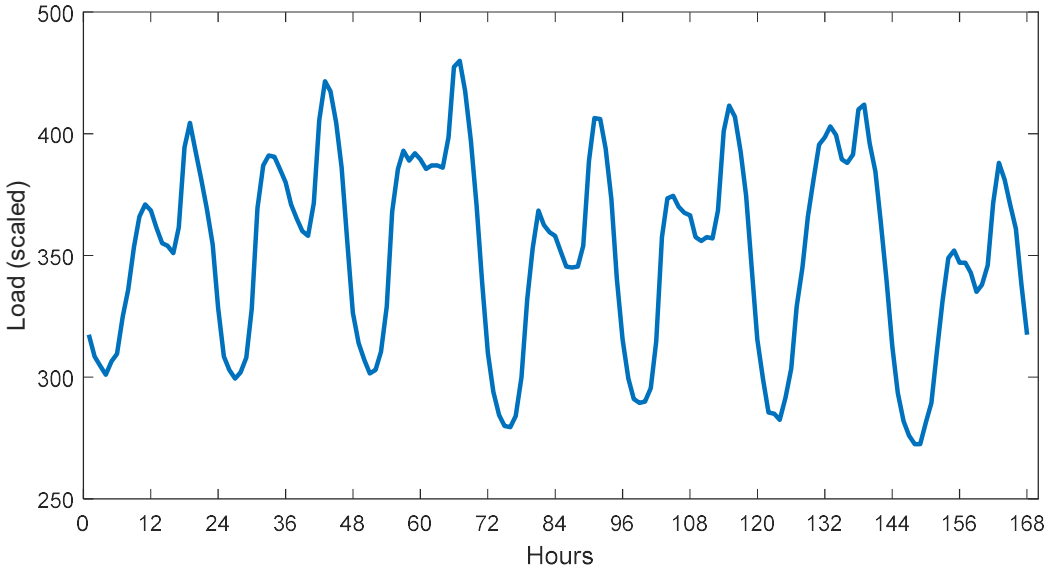

Figure 3.2: Weekly energy load patterns.

Table 3.1: Comparison of point forecasting model factors.

	MLR1 [46]	MLR2 [78]	MLR3 (new)
Weather inputs	Temperature Wind speed Dew point Precipitation	Temperature Dry Bulb Dew point	Temperature Wind speed Dew point
Seasonality	Hour Weekday Day of the year	Hour Weekday Holiday (12AM -11PM)	Hour Weekday Holiday (8AM – 6PM) Day of the year
Historical load	1, 2, 4, 8, 12, 24, 48, 72, and 168 hours ago.	1, 24, and 168 hours ago.	1-3, 23-25, and 167-169 hours ago.
Historical weather	1, 2, 4, 8, 12, 24, 48, 72, and 168 hours ago.	None.	1 and 2 hours ago.

Denoting temperature as T, wind speed as WS, dew point as DP, load (flow) as Y, day of the year as DOY, day of the week as DOW, and hour of the day as HOD, the MLR3 method is

$$\begin{aligned}\hat{Y}_k = \quad & \beta_0 + \sum_{i=1}^{2}\left(\beta_i * T_{k-i} + \beta_{i+2} * WS_{k-i} + \beta_{i+4} * DP_{k-i}\right) \\ & + \sum_{i=1}^{3}\left(\beta_{i+6} * Y_{k-i} + \beta_{i+9} * Y_{k-22-i} + \beta_{i+12} * Y_{k-166-i}\right) \\ & + \sum_{i=1}^{2}\left(\beta_{14+2i} * \sin\left(2\pi i * \frac{DOY}{365}\right) + \beta_{15+2i} * \cos\left(2\pi i * \frac{DOY}{365}\right)\right) \\ & + \sum_{i=1}^{2}\left(\beta_{18+2i} * \sin\left(2\pi i * \frac{DOW}{7}\right) + \beta_{19+2i} * \cos\left(2\pi i * \frac{DOW}{7}\right)\right) \\ & + \sum_{i=1}^{2}\left(\beta_{22+2i} * \sin\left(2\pi i * \frac{HOD}{24}\right) + \beta_{23+2i} * \cos\left(2\pi i * \frac{HOD}{24}\right)\right).\end{aligned} \tag{3.1}$$

The MLR3 model is used to generate point forecasts for all three probabilistic forecasting methods described in the next section: a) a benchmark model based on a normality assumption (NDEPF), b) kernel density estimation based probabilistic forecasts (KDEPF), and c) probabilistic forecasts using data transformation by a Johnson curve (JDTPF). Each of these methods has three variants.

3.2 Probabilistic Forecasts using a Normality Assumption - A Benchmark

This section presents the first of three probabilistic forecasting methods presented in this dissertation, the normal density estimator probabilistic forecasting (NDEPF) method. The NDEPF method assumes that the forecasting error is normal, although the forecasting error distribution is typically not normal in practice (see Figure 3.3). Many probabilistic forecasting articles discussed in Section 2.2 assume normality of the error distribution to calculate probabilistic forecasts. Moreover, Tom Connery and Tyler Stephens in their natural gas industry overview talk in GasDay™ [2] mentioned the use of a normal distribution to quantify forecast uncertainty as a temporary solution of their problem (see Section 1.1) because of the easy calculation process. Thus, the probabilistic forecasts assuming normality, NDEPF is considered as a benchmark model in this dissertation.

The MLR3 method presented in Section 3.1 is used to generate point forecasts from historical weather and energy demand. Then, hourly historical forecasting residuals (r) are calculated from actual energy demand (Y) and point forecasts ($\hat{Y}$),

$$r = Y - \hat{Y} . \tag{3.2}$$

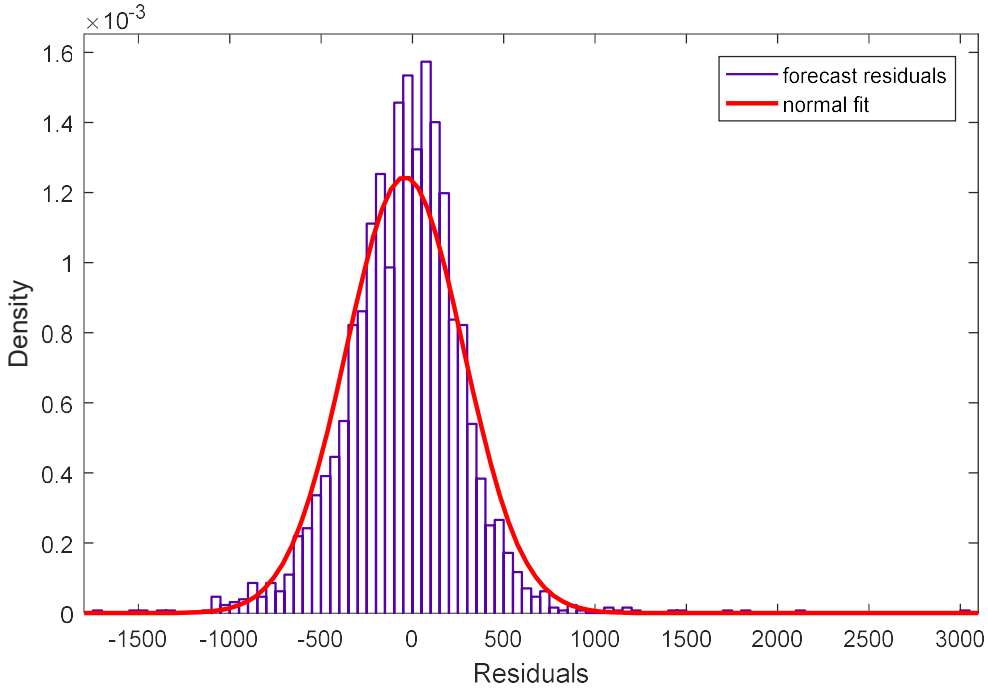

Figure 3.3: A sample error distribution from real data.

Now, residuals are sorted by conditionals to make several small groups of residuals. Three conditionals have been used in this work: 1) temperature (T), 2) last 24-hour temperature difference

$$\Delta T_k = T_k - T_{k-24}, \tag{3.3}$$

and 3) difference between current temperature and last 168 hours (1 week) average temperature

$$\Delta T_k\big|_{wk} = T_k - \frac{\sum_{i=k-169}^{k-1} T_i}{168}. \tag{3.4}$$

Finally, residuals are divided into small subsets called residual bins based on one of the three conditionals. Other conditionals, such as heating degree day (HDD), cooling degree day (CDD), or multiple conditionals could be used (see Section 5.3). The flow chart of the residual binning process of the NDEPF method is shown in Figure 3.5. Each residual bin must contain enough residuals to create an error distribution. Thus, a minimum bin size (number of residuals) is enforced. Intentional overlapping of 20 percent residuals between two consecutive residual bins avoids discontinuities (see Figure 3.8). Each residual bin has three properties: a) start and end index, which represent the range of the conditional, b) number of residual samples, and c) a residual cumulative distribution function (CDF). In this method, residual CDFs are estimated using a parametric approach - normal distribution. Figure 3.4 shows the overall view of generating probabilistic forecasts for the NDEPF method. Here, the first flow chart (Figure 3.5) is analogous to training, and the second flow chart (Figure 3.7) is similar to evaluation.

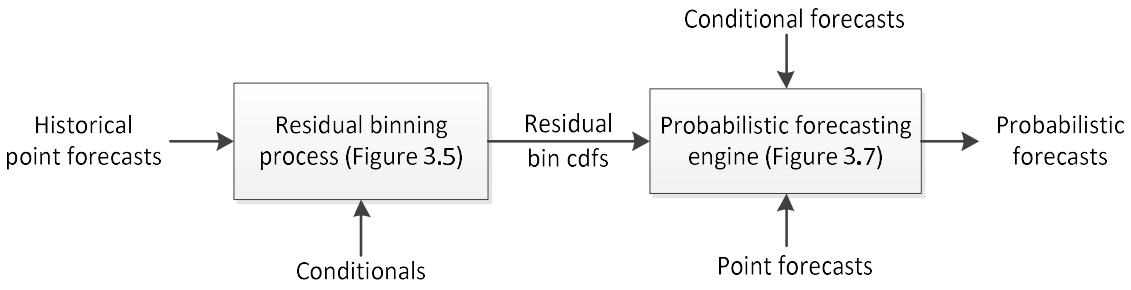

Figure 3.4: Overview of the probabilistic benchmark method flowcharts.

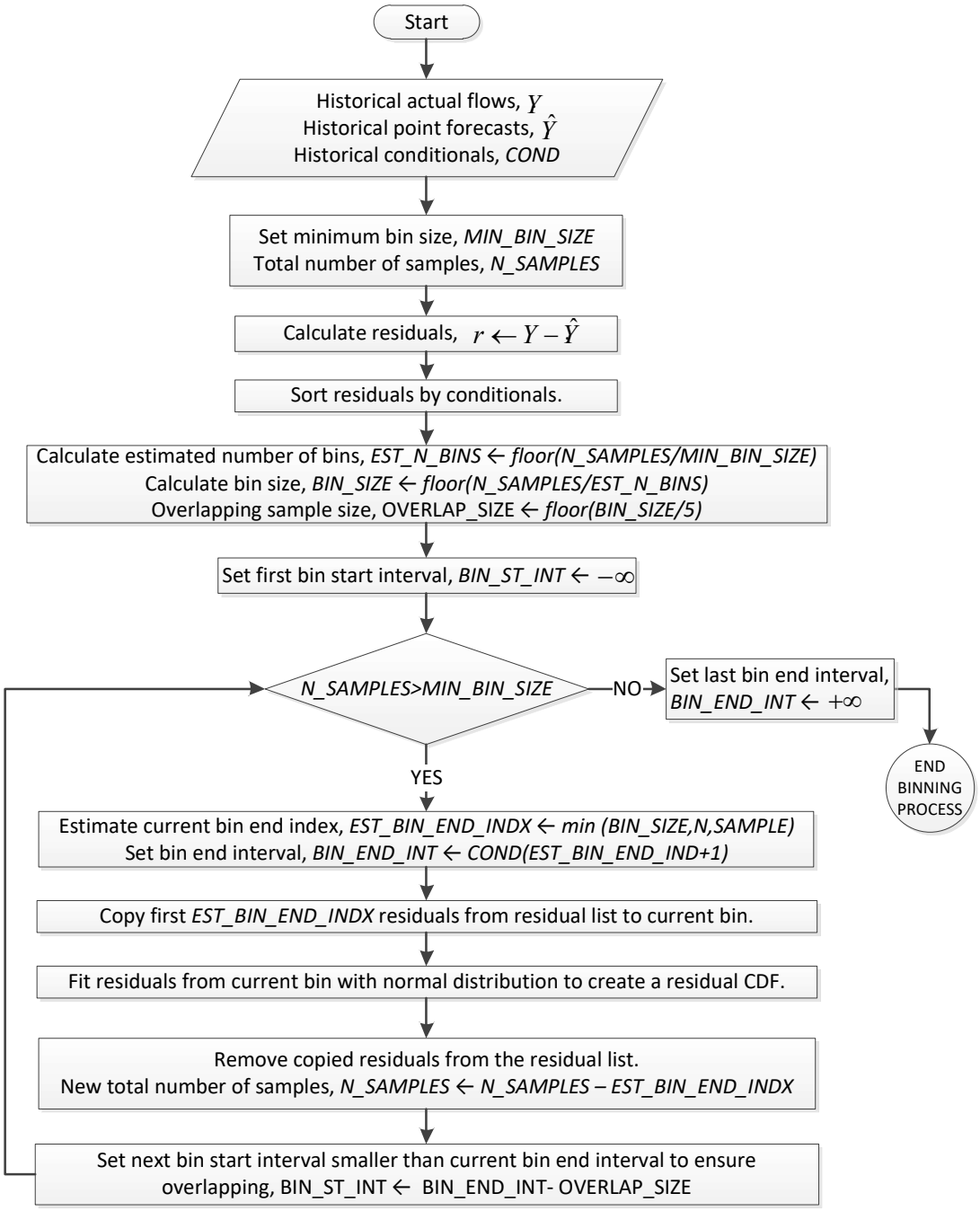

Figure 3.5: Flow chart of residual binning process assuming normality (a benchmark).

The normal distribution [190] has some unique properties, which make it possible to calculate the entire distribution by knowing only its mean (μ) and standard deviation (σ). The normal probability distribution function (PDF) is

$$f\left(r \mid \mu, \sigma\right) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(r-\mu)^2}{2\sigma^2}} . \tag{3.5}$$

If the error distribution fits a normal distribution, 68% of the data are within one standard deviation of the mean, 95% of the data are within two standard deviations, and 99.7% of the data are within three standard deviations of the mean [191]. If *r* is a data point (residual) in a normal distribution, it can be converted into a standard score (z-score),

$$z = \frac{r - \mu}{\sigma} . \tag{3.6}$$

The z-score is easily convertible to a percentile (1% to 99%) from an inverse CDF using the property of normal distributions shown in Figure 3.6 (see also standard statistical textbooks such as [192, 193, 194] for more details). This process is repeated for each of the residual bins to get percentiles from inverse residual CDFs for the NDEPF method.

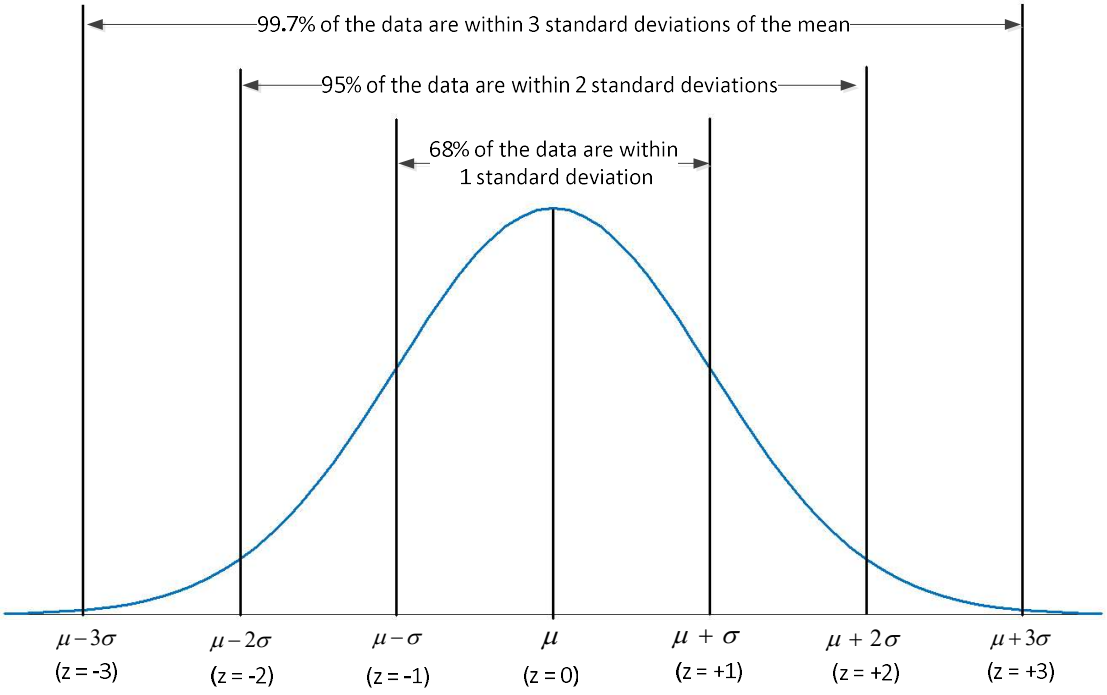

Figure 3.6: A sample normal distribution curve.

Once the residual CDFs are made within each residual bin, the probabilistic forecasting engine is ready to generate a probabilistic forecast. A new probabilistic forecast is made based on a point forecast and forecasted conditional. The forecasted conditional is used to select a residual bin, whose residual CDF is retrieved. Then a point forecast is added to each quantile (1% to 99%) of the residual CDF to generate a probabilistic forecast. The mean of the probabilistic forecasts (forecasted CDF) has shifted (mean = point forecast + μ) from the residual CDF mean (μ), the variance (σ) does not change. Figure 3.7 shows the flow chart of generating a probabilistic forecasting from an existing point forecast, forecasted conditional, and residual bins.

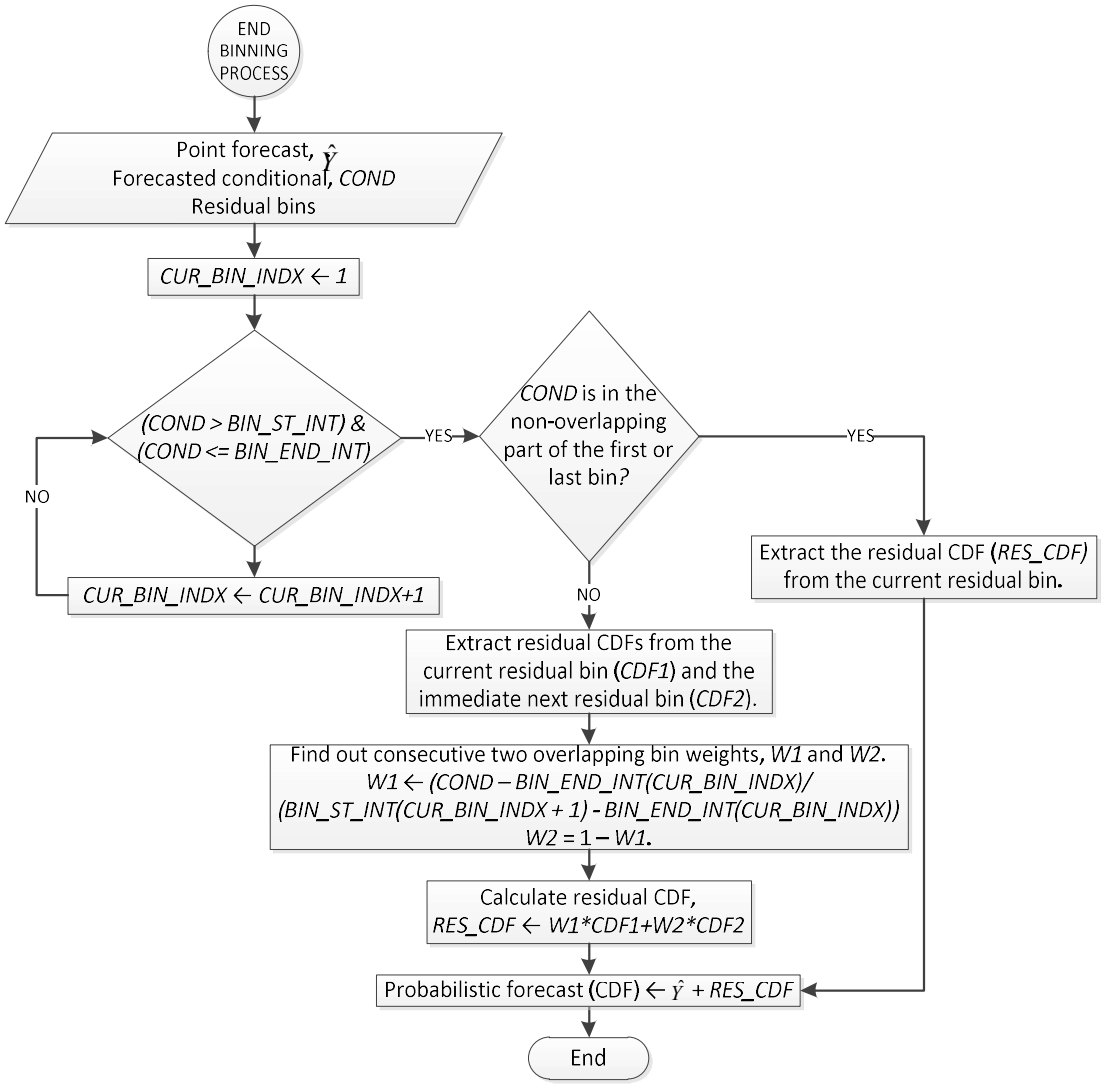

Figure 3.7: Flow chart of generating probabilistic forecasts.

An illustrative example of the binning process is shown in Figure 3.8. Here, temperatures (in Fahrenheit) are considered as the conditionals. Suppose, only 13 samples were available in our cartoon training dataset (see Figure 3.8). In practice, there are far more than 13 samples in a training dataset. For example, the experimental result presented in Chapter 4 contains approximately 43,800 samples (5 years of hourly data) in the training dataset. Residuals are calculated from point forecasts using Equation (3.2).

Then, the training dataset is sorted by temperature as shown in Figure 3.8. The residual bins are created and populated with residuals according to the flow chart of Figure 3.5. In the Figure 3.8 scenario, there are only three residual bins. However, more than 20 bins are created when the algorithm is used on a real dataset. Similarly, more than one sample is shared in consecutive residual bins. Then, residual CDFs are constructed applying a normal fit on all residual samples within each residual bin.

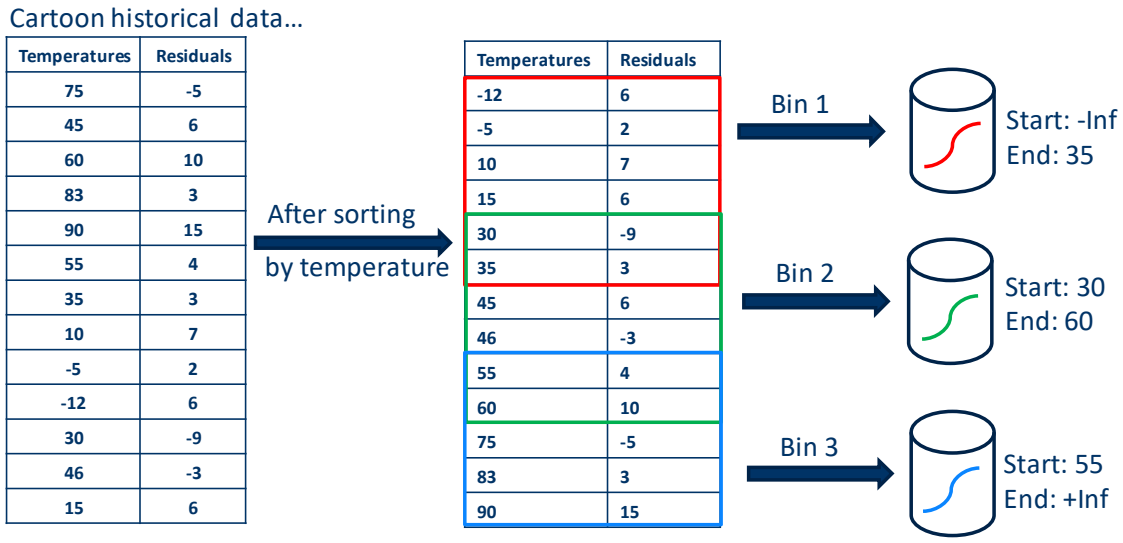

Figure 3.8: An example of the binning process using a cartoon dataset.

Now, the residual bins are ready to use for generating probabilistic forecasts. The real testing datasets (used for the experiment in Chapter 4) contain 17,520 samples (2 years of hourly data). However, only three samples are shown in the cartoon testing dataset (see Figure 3.9). Consider, the first row of the cartoon testing dataset, which contains a forecasted conditional (32°F) and a point forecast (650 units). The forecasted temperature, 32°F (conditional), is used to find the correct residual bins. A conditional

may fit with two consecutive residual bins because of overlapping 20 percent of the residuals between two consecutive bins. In this case, the first (range: -Inf to 35°F) and second (range: 30°F to 60°F) residual bins are the desired residual bins, because 32°F fits between the range of these two overlapping bins. The final residual CDF is the weighted combination of these two residual CDFs as shown in the inset of Figure 3.9. Finally, the mean of the resultant residual CDF is shifted by the point forecast (i.e., adding point forecast with each of the 99 quantiles of the residual CDF) to generate a probabilistic forecast for the given temperature (conditional). This entire process is repeated for each time horizon.

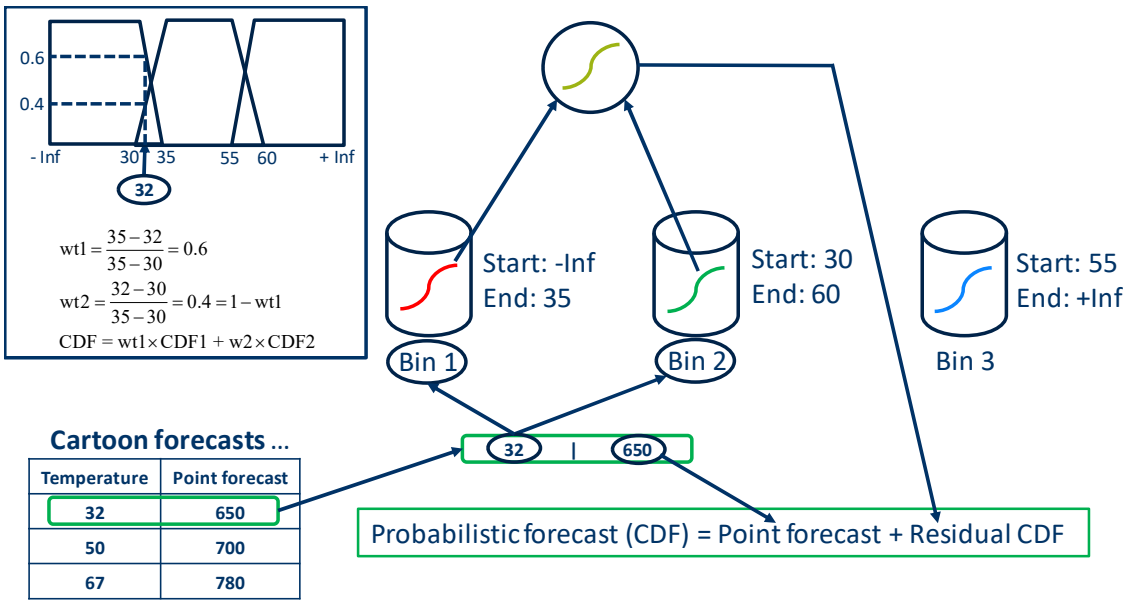

Figure 3.9: An example of generating probabilistic forecasts from cartoon forecasts.

Other possible conditionals include the last 24 hours temperature difference (see Equation (3.3)) and the difference between the current temperature and the last 168 hours

average temperature (see Equation (3.4)). Probabilistic forecasts generated by both of these conditionals are comparable with the temperature conditional (see Chapter 4). Multiple conditionals could be used (see Section 5.3).

The next two sections present two new probabilistic forecasting methods without assuming error normality. The first one is based on a non-parametric approach called kernel density estimator (KDE). The second one is based on a data transformation technique called the Johnson curve [188, 195]. Both KDE probabilistic forecasting (KDEPF) method and Johnson data transformation probabilistic forecasting (JDTPF) method use the same binning process and probabilistic forecasting engine, respectively, as the benchmark probabilistic forecasting method, NDEPF.

3.3 Probabilistic Forecasts Using a Kernel Density Estimator

This section presents the second of three major probabilistic forecasting methods, kernel density estimator probabilistic forecasting (KDEPF). This method also creates residual bins based on conditionals such as 1) temperature, 2) daily temperature difference (see Equation (3.3)), or 3) difference between current temperature and last 168 hour average temperature (see Equation (3.4)) similar to the benchmark method, NDEPF (see Figure 3.8). However, the residual CDFs for each of the residual bins are estimated using a non-parametric approach, a kernel density estimator (KDE) [196, 197, 198]. Figure 3.10 illustrates the flow chart of the residual binning and training CDF process using KDE. The residual binning part of the flow chart is similar to the NDEPF binning

flow chart, but the training CDF part is different. However, it may be helpful to see the entire process at a glance.

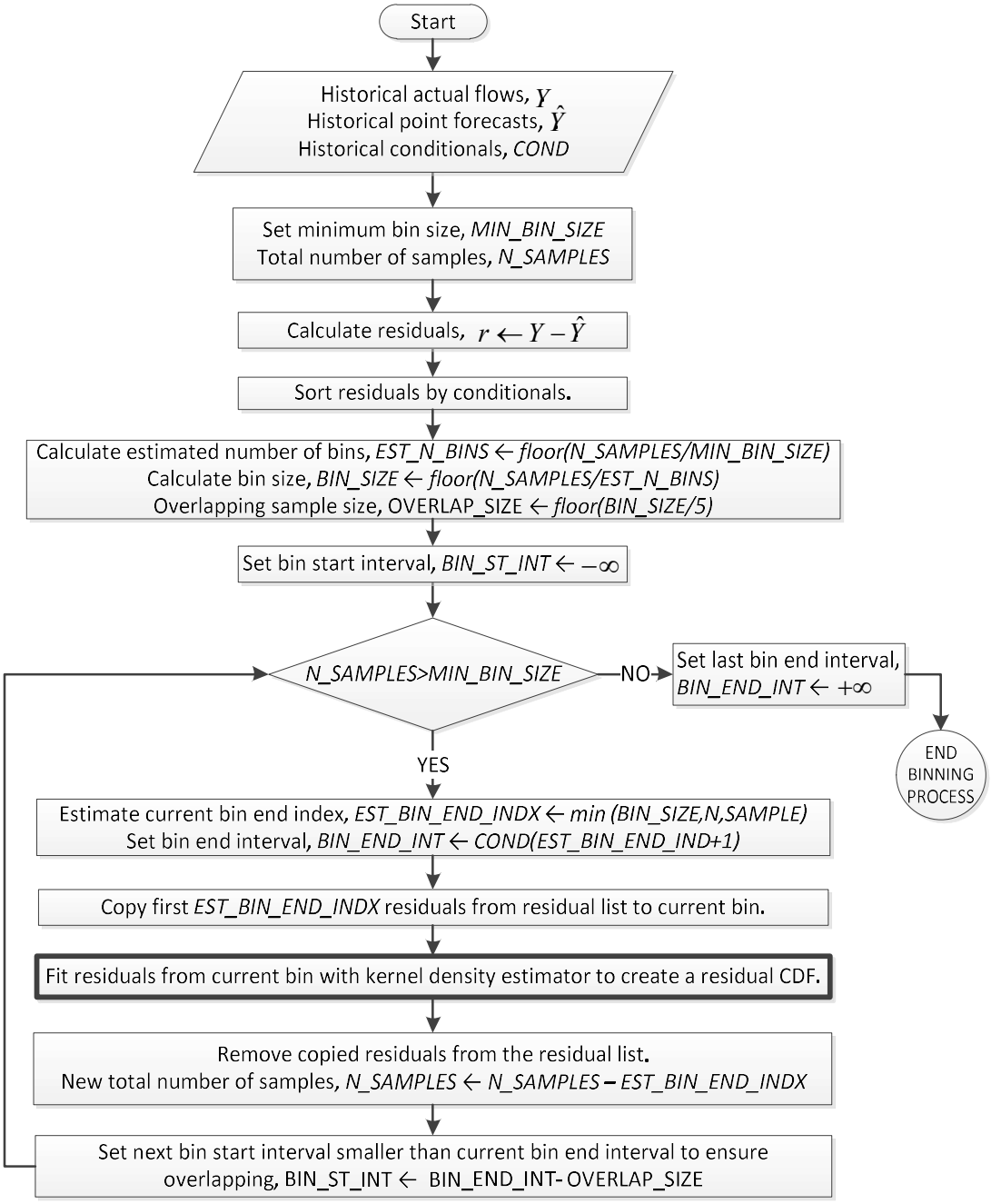

Figure 3.10: Flow chart of the residual binning process using KDE.

A kernel distribution is a nonparametric distribution of a random variable. It is used when a parametric distribution poorly represents the dataset. In this dissertation, an initial attempt to fit forecasting error with known distributions, such as normal, beta, gamma, binomial, logistic, exponential, Weibull, and Rayleigh did not work well. Thus, a KDE is used to describe forecasting error. If n is the sample size, $K(.)$ is the kernel smoothing function, and the bandwidth is denoted by h, then the KDE [196, 197] is

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \quad ; -\infty < x < \infty \tag{3.7}$$

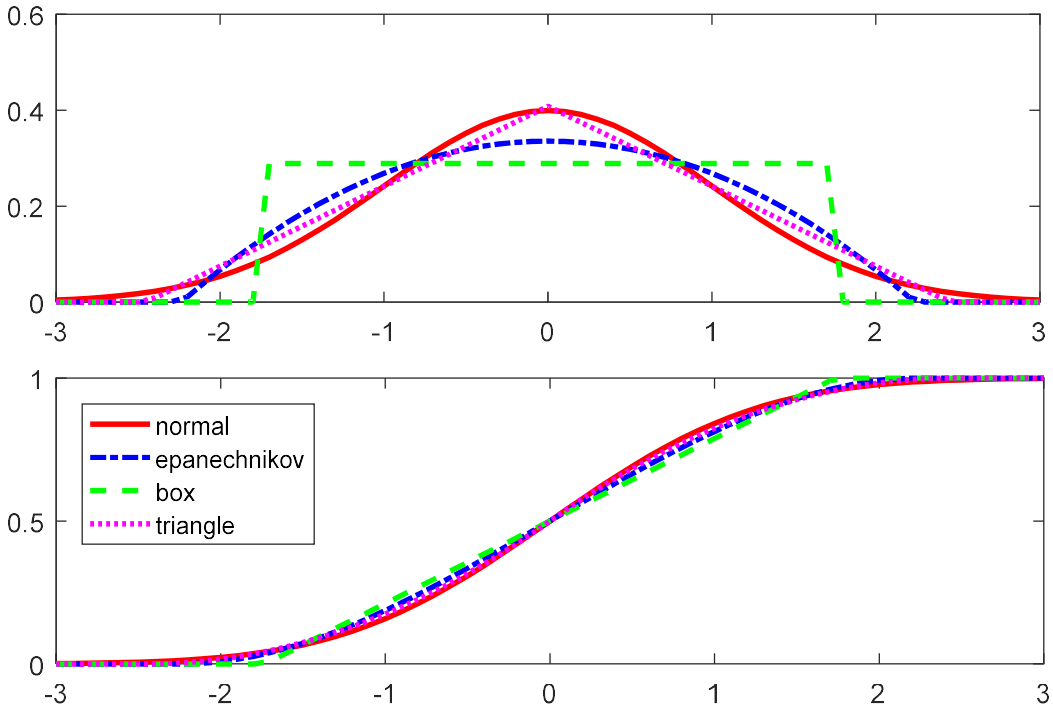

Figure 3.11: Different smoothing functions in used in KDE (adapted from [196]).

Various smoothing functions used in KDE [196, 198] (such as normal (Gaussian), Epanechnikov, uniform (box), and, triangle) define the shape of the curve used to generate the CDF or PDF (see Figure 3.11). In this work, CDFs are used to represent probabilistic forecasts. However, the difference between various smoothing functions are more visible as PDFs. Thus, both PDF and CDF are provided for better understanding. The normal smoothing function is used in this work. A sample kernel distribution based on only six cartoon data samples using a normal smoothing function is shown in Figure 3.12. Both PDF and CDF version of the graph are provided for better understanding.

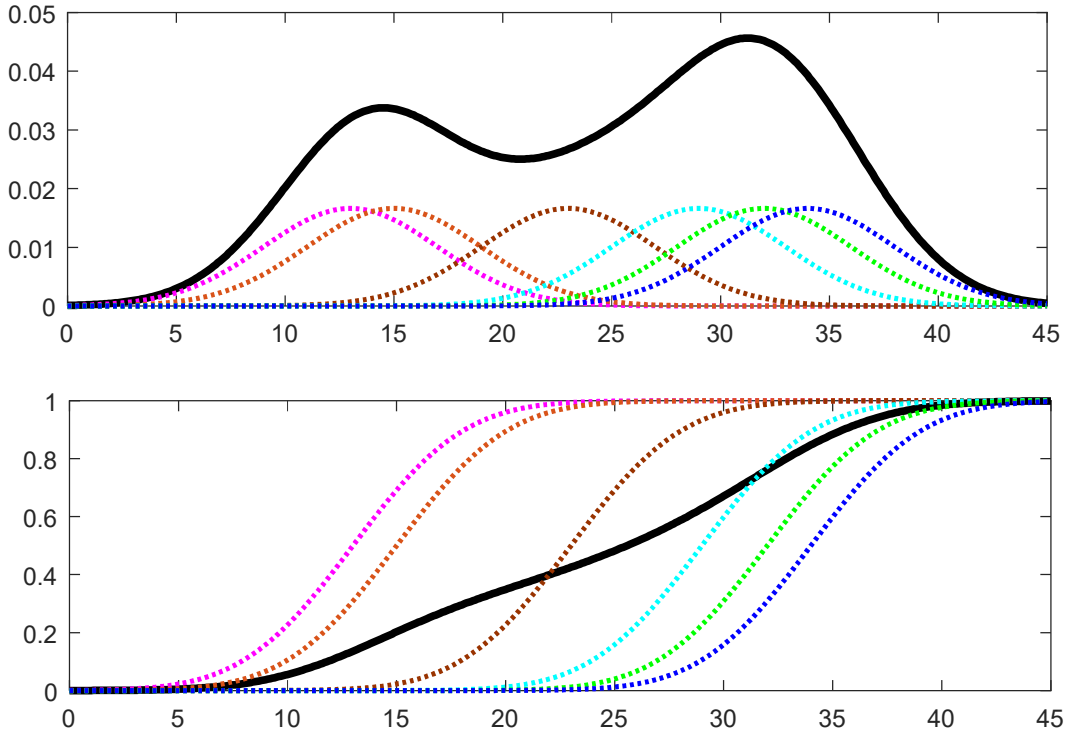

Figure 3.12: A sample kernel distribution using cartoon dataset (adapted from [196]).

The bandwidth controls the smoothness of the PDF. If it is too small, then it generates a rough curve, which provides tiny details of the PDF. On the other hand, if the bandwidth is too large, then some important features of the dataset might be obscured. Figure 3.13 shows the effect of using different bandwidths on a cartoon dataset containing only six samples.

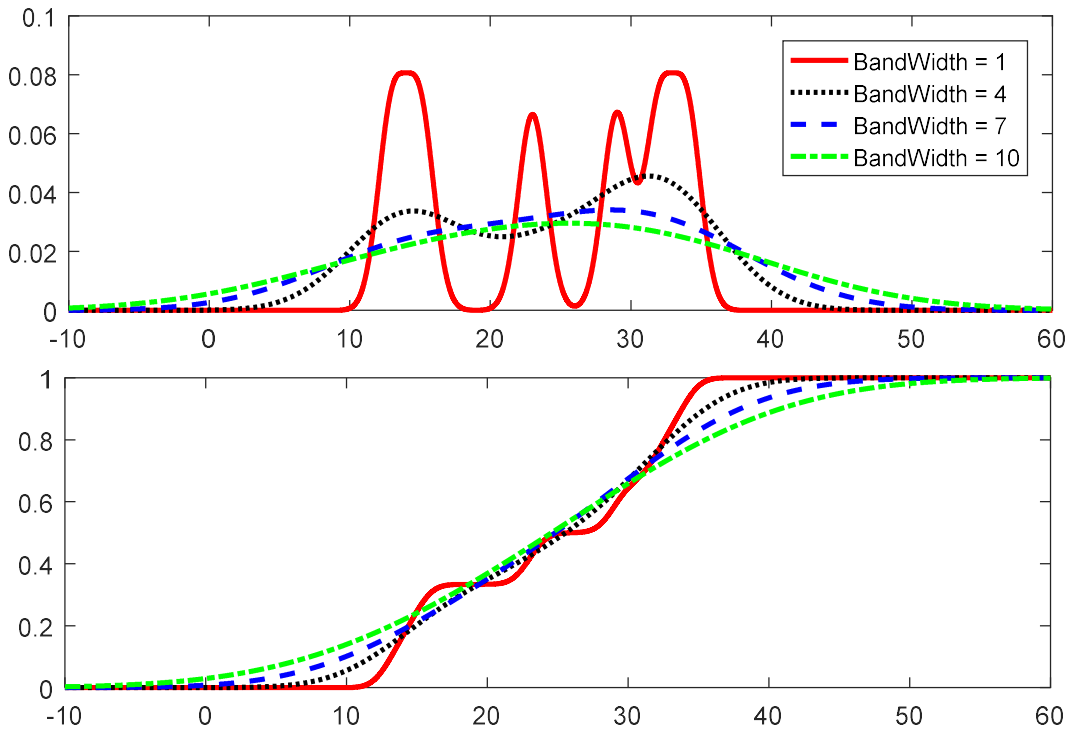

Figure 3.13: Bandwidth selection for KDE (adapted from [196]).

The default bandwidth in the MATLAB statistics and machine learning toolbox estimates a theoretically optimal density function for the normal kernel smoothing function, which produces a reasonably smooth curve. In this work, the default bandwidth with the normal kernel smoothing function was selected. Figure 3.14 shows a residual

CDF and its corresponding PDF generated from a sample residual bin using a real dataset.

Once the residual CDFs are generated from the historical dataset, probabilistic forecasts can be generated from point forecasts and forecasted conditionals. The process of generating probabilistic forecasts using KDE is similar to the benchmark probabilistic forecasting method (NDEPF), explained in the previous section. A flow chart of the probabilistic forecasting engine is shown in Figure 3.7.

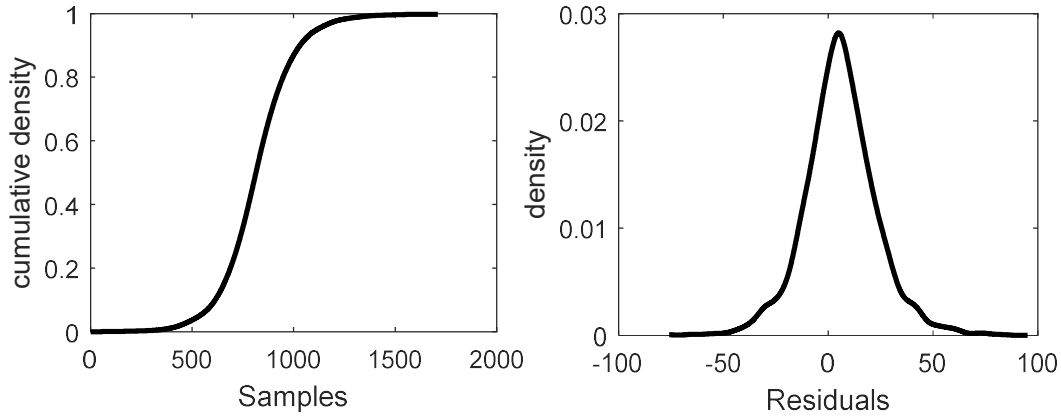

Figure 3.14: A sample residual CDF and PDF calculated from a residual bin using KDE.

According to the case studies in Chapter 4, the KDEPF method performed better than the benchmark NDEPF method, based on scores provided by all four scoring rules (Pinball, CRPS, QCS, and PQCS). However, the KDEPF method requires about 100 times the CPU time of the benchmark NDEPF (see Section 4.2.4). The next section proposes a new faster way (as fast as NDEPF) to calculate residual CDF without the

normality assumption, and whose scores (Pinball, CRPS, QCS, and PQCS) are comparable with the KDEPF.

3.4 Probabilistic Forecasts Using the Johnson Data Transformation

This section proposes a new way to generate probabilistic forecasts using the Johnson data transformation technique [188, 195], called the Johnson data transformation probabilistic forecasting (JDTPF) method. The residual binning process of the JDTPF method is similar to the other two probabilistic forecasting methods, NDEPF (see Section 3.2) and KDEPF (see Section 3.3), presented in this dissertation. An illustrative example of the residual binning process is explained with a cartoon example in Section 3.2 (see Figure 3.8). Once the residual bins are created based on conditionals (temperature, last 24- hour temperature difference, or difference between current temperature and last week's average temperature), the Johnson transformation can be used to transform non-normally distributed errors into nearly normally distributed data. Figure 3.15 illustrates a high-level work flow diagram of the Johnson Curve Toolbox in MATLAB [199] used in this work.

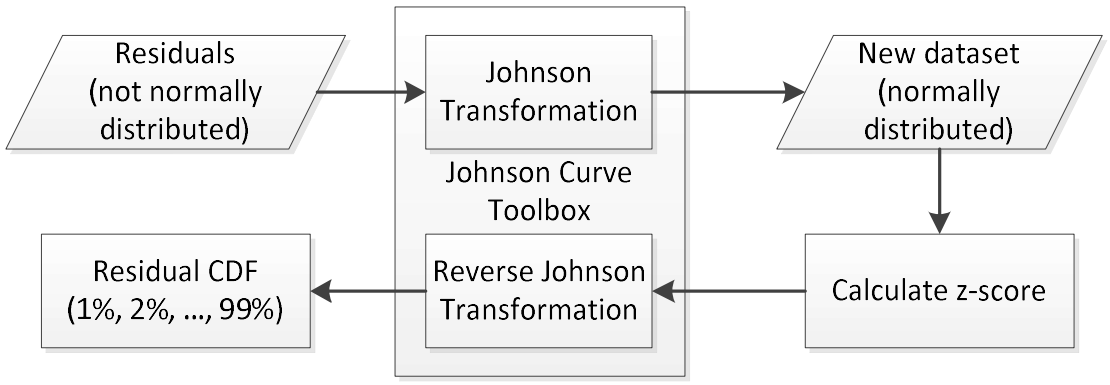

Figure 3.15: Use of the Johnson Curve Toolbox to calculate residual CDFs in MATLAB.

The Johnson Curve is a very powerful and flexible tool to transform non-normal distributions into normal distributions based on three families of transformations: 1) exponential, 2) logistic, and 3) hyperbolic sine shown in Equations (3.8), (3.9), and (3.10), respectively [188, 195, 200]. The Johnson system of distributions (Johnson Curves) can be defined by four parameters $(\gamma, \delta, \xi, \lambda)$. The first two parameters (γ, δ) define the shape of the distribution (like skewness and kurtosis), the third parameter (ξ) denotes the location of the median, and the fourth parameter (λ) indicates the scale of the distribution (similar to standard deviation). So, the job of the 'Johnson Transformation' black box mentioned in Figure 3.15 is to find four transformation parameters $(\gamma, \delta, \xi, \lambda)$ from a given non-normal dataset. In this work, the unbounded Johnson distribution function was selected because the maximum (or minimum) forecast error is not known in advance.

Log-normal (SL): $$z = \gamma + \delta \ln\left[\frac{(x-\xi)}{(\lambda+\xi-\chi)}\right]. \tag{3.8}$$

Unbounded (SU): $$z = \gamma + \delta \ln\left(\frac{x-\xi}{\lambda}\right). \tag{3.9}$$

Bounded (SB): $$z = \gamma + \delta \sinh^{-1}\left[\frac{(x-\xi)}{\lambda}\right]; \quad \text{where, } \sinh^{-1}(x) = \ln\left[x + \sqrt{(1+x^2)}\right]. \tag{3.10}$$

Johnson's (1949) original procedure [188] for finding the four transformation coefficients is based on moments derived from the given dataset. In 1952, Draper improved the accuracy of the calculation by suggesting an algebraic formula replacing the original graphical calculation technique [201]. Wheeler proposed an alternative method of fitting a Johnson distribution to data based on quantiles instead of moments in 1980 [202]. Both methods (moments and quantiles) of fitting Johnson distribution are available in the Johnson Curve Toolbox in MATLAB [199] (an earlier version was written in FORTRAN [203, 204]). The quantile method is used in this work. Figure 3.16 shows the performance of the Johnson transformation on a real dataset (forecast errors). Other normality tests such as Jarque-Bera, Lilliefors, and Kolmogorov-Smirnov (with 5% significance level) are provided also in Table 3.2.

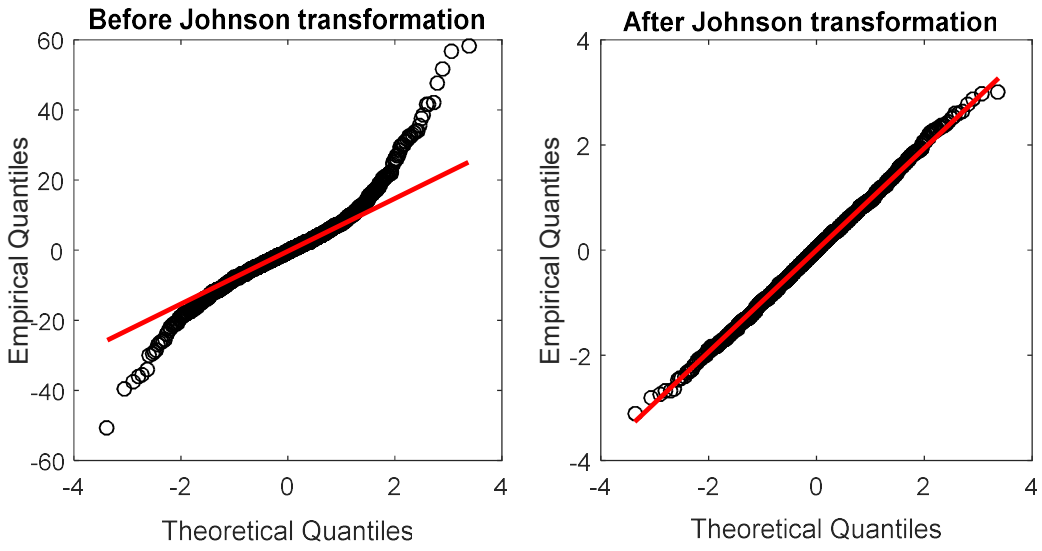

Figure 3.16: Normality check of Johnson transformation using qq-plots.

Table 3.2: Normality tests of a Johnson transformation.

Normality Tests (5% significance level)	Original Data		Transformed Data	
	H_0: normal H_a: non-normal	p-value	H_0: normal H_a: non-normal	p-value
Jarque-Bera test	Reject	0.1606	Accept	0.0010
Lilliefors test	Reject	0.5000	Accept	0.0010
Kolmogorov-Smirnov test	Reject	0.8188	Accept	0.0000

The Box-Cox transformation is another well-known data transformation technique [205] used for a similar purpose as the Johnson Curve. The Box-Cox transformation is easier to understand compared to the Johnson transformation. However, it does not work for zero and negative values [206]. On the other hand, the Johnson transformation is powerful, flexible, and can work with data including zero and negative values. In this dissertation, the data transformation technique is applied to forecasting errors, which are

expected to be negative or zero approximately 50% of the time. Thus, the Johnson transformation has been chosen over the Box-Cox transformation.

The ‘Johnson Transformation’ black box shown in Figure 3.15 transforms a non-normal data into an approximately normal data (see Figure 3.16 and Table 3.2 as an example). The next paragraph provides more information about the ‘Johnson Transformation’ black box. The z-scores are calculated from the transformed normal distribution (see Figure 3.6). The ‘Reverse Johnson Transformation’ black box does the reverse transformation of what the ‘Johnson Transformation’ does. Here, the reverse Johnson transformation is applied to calculated z-scores to construct the required residual CDFs. The residual binning process of the JDTPF method is similar to the NDEPF and KDEPF, but the learning CDF part is different (shown in Figure 3.15). However, the full flowchart of the residual binning process for the JDTPF method is shown in Figure 3.17 for a better understanding of the whole scenario.

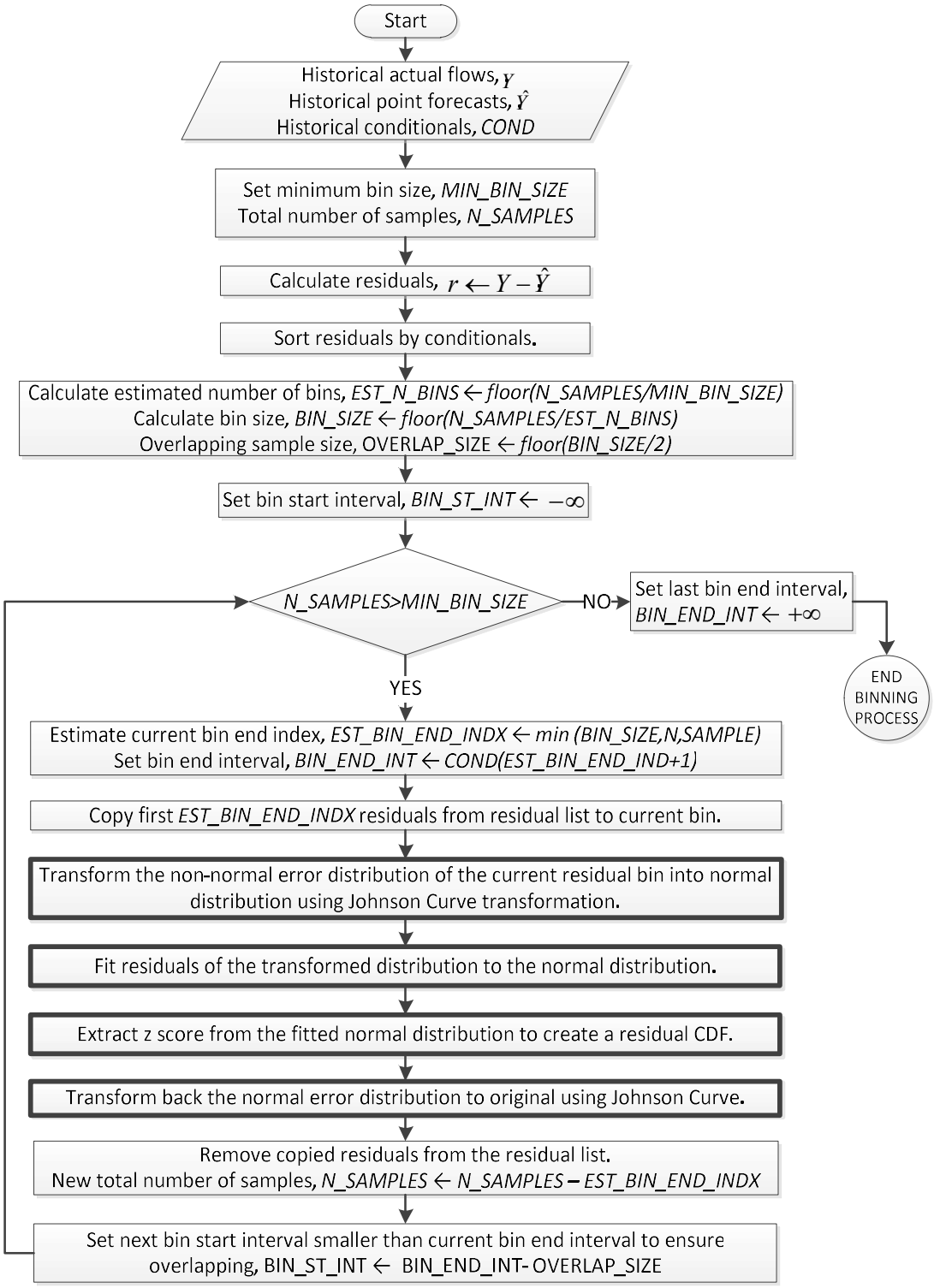

Figure 3.17: Flow chart of the residual binning process using the Johnson data transformation.

The performance of all three probabilistic forecasting methods (NDEPF, KDEPF, and JDTPF) is assessed and compared in Section 4.2 using two new scoring rules, QCS and PQCS, presented in the next section. A new evaluation technique is required because existing evaluation techniques (see Section 2.3) are complicated and focus more on sharpness (see Figure 2.2), which is less important than calibration (see Figure 2.1) in practice.

3.5 A New Evaluation Technique for Probabilistic Forecasts

This section presents the second major contribution of this dissertation. Forecast evaluation techniques are required to assess the effectiveness of forecasts generated by probabilistic forecasting engines. Evaluating probabilistic forecasts is as important as generating probabilistic forecasts because a good evaluation technique guides researcher to produce useful probabilistic forecasts. Hong asserts that, a lack of well-established evaluation techniques is one reason for the underdevelopment of probabilistic forecasting research [37].

The concept of a probabilistic forecast is more difficult to understand than the concept of a point forecast because the probabilistic forecast presents a complete distribution of a future event compared to only one value (50^{th} percentile) offered by a point forecast. Hence, a visual representation of a probabilistic forecast is helpful. Similarly, the evaluation of a probabilistic forecast should contain two parts (graphical and numerical) to make the probabilistic forecast evaluation easier to comprehend. Based

on an extensive literature search, no evaluation technique that presents evaluation results in both graphical and numerical formats is known. The new evaluation technique proposed in this section has both a graphical representation, graphical calibration measure (GCM), and two numerical scores, quantile calibration score (QCS) and percentage quantile calibration score (PQCS).

The most popular graphical technique used for evaluating probabilistic forecasts is the probability integral transform (PIT) [124]. The PIT is useful to obtain a rough idea of the calibration or reliability (see Figure 2.1, subsection 2.2.1) of a probabilistic forecast. However, it is difficult to compare two probabilistic forecasting methods using PIT applied on the same dataset because they look almost identical. A numeric evaluation technique is needed.

Several scoring rules used to assess probabilistic forecasts are explained in Section 2.3. Reliability (see Figure 2.1), sharpness (see Figure 2.2), and resolution (variation of the forecast CDF with time) are considered three criteria of a good probabilistic forecast [37]. The continuous ranked probability score (CRPS) and the pinball loss function are the most used evaluation techniques in the recent probabilistic load forecasting literature. Hence, these two scoring rules are included as base-line scoring rules to assess probabilistic forecasts with our two scoring rules (QCS and PQCS).

Landry et al. [164] showed an obvious weakness of the pinball score. Landry's probabilistic forecasting method secured the first place in the wind power forecasting

track of GEFCom2014 [112] based on the pinball loss function evaluation technique. However, the percentage observation column in the Table 3.3 reveals some weaknesses of the produced probabilistic forecast (shown in bold). Figure 3.18 shows the graphical version of Table 3.3 (which is like our graphical evaluation technique, GCM). The result (in Table 3.3 and Figure 3.18) shows that the probabilistic forecast is left-skewed and concentrated in the middle of the distribution (59.5% of the observations are within the 20^{th} and 50^{th} quantiles, where the expected observation should be 30%). Again, very few observations are in the right tail of the distribution (only 7% of the observations are between the 60^{th} and 100^{th} quantiles, while 40% was expected). Overall, the probabilistic forecast is not well-calibrated. The pinball loss function is biased toward sharp probabilistic forecasts. Both the too sharp and less sharp are not useful probabilistic forecast in practice.

Table 3.3: Pinball score and percentage observation of Landry's wind power probabilistic forecasts for the GEFCom2014 (adapted from [164]).

Quantile range	Mean pinball loss	% observation	% expected
0.0-0.1	0.0039	6.6%	10%
0.1-0.2	0.0143	13.9%	10%
0.2-0.3	0.0265	**18.5%**	10%
0.3-0.4	0.0364	**21.8%**	10%
0.4-0.5	0.0512	**19.2%**	10%
0.5-0.6	0.0597	10.4%	10%
0.6-0.7	0.0628	**6.2%**	10%
0.7-0.8	0.0538	**0.7%**	10%
0.8-0.9	0.0676	**0.1%**	10%
0.9-1.0	-	**0.0%**	10%

A useful probabilistic forecast is well-calibrated, i.e., the percentage of the total observed values is close to the percentage expected. CRPS and other scoring rules have similar problems, giving too much importance to high sharpness. In addition, there is no graphical version of these scoring rules, which makes them more difficult to understand. Our evaluation technique, graphical calibration measure (GCM), has a graphical version with two numerical scores (QCS and PQCS), which makes it easier to understand. It provides a better score for being close to the percentage expected and penalties for being too sharp or less sharp.

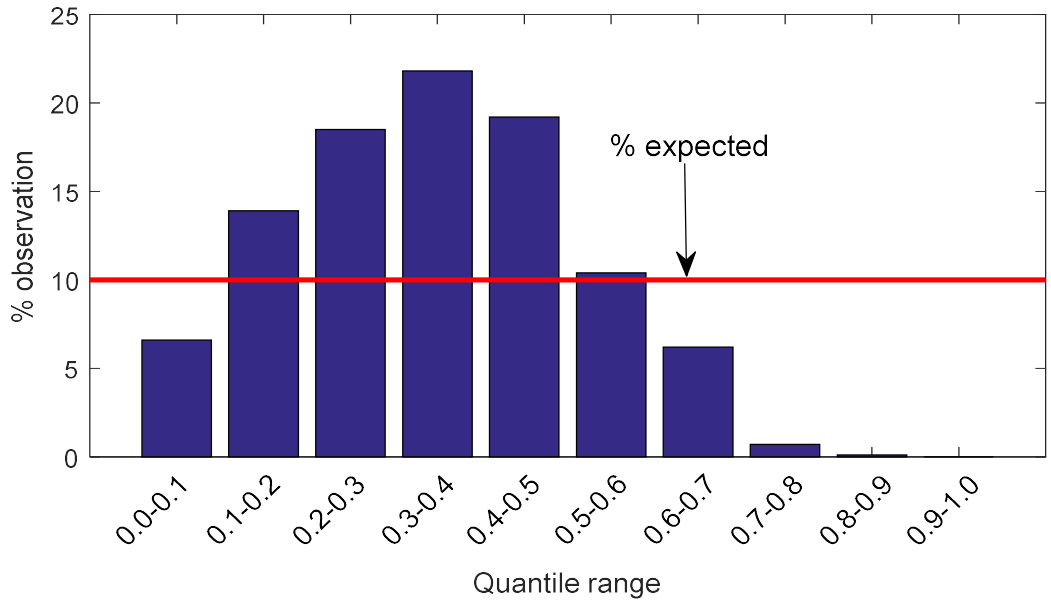

Figure 3.18: Performance of the Landry's probabilistic wind power forecasting model [164].

The flow chart of the graphical calibration measure (GCM) is shown in Figure 3.19. The GCM divides a forecast CDF into several quantile bins for assessment. The total number of quantile bins is calculated from 100 divided by bin width (a user input

between 1 and 100). The default bin width (used in this work) is 10. Hence, the default number of quantile bins is 10. The main idea of the GCM is to fill up all quantile bins with available observed values and then check the deviation of the quantile bin population from expectation. For a well-calibrated probabilistic forecast, the expected population (actual values) inside each bin should be approximately the total observed values divided by bin width.

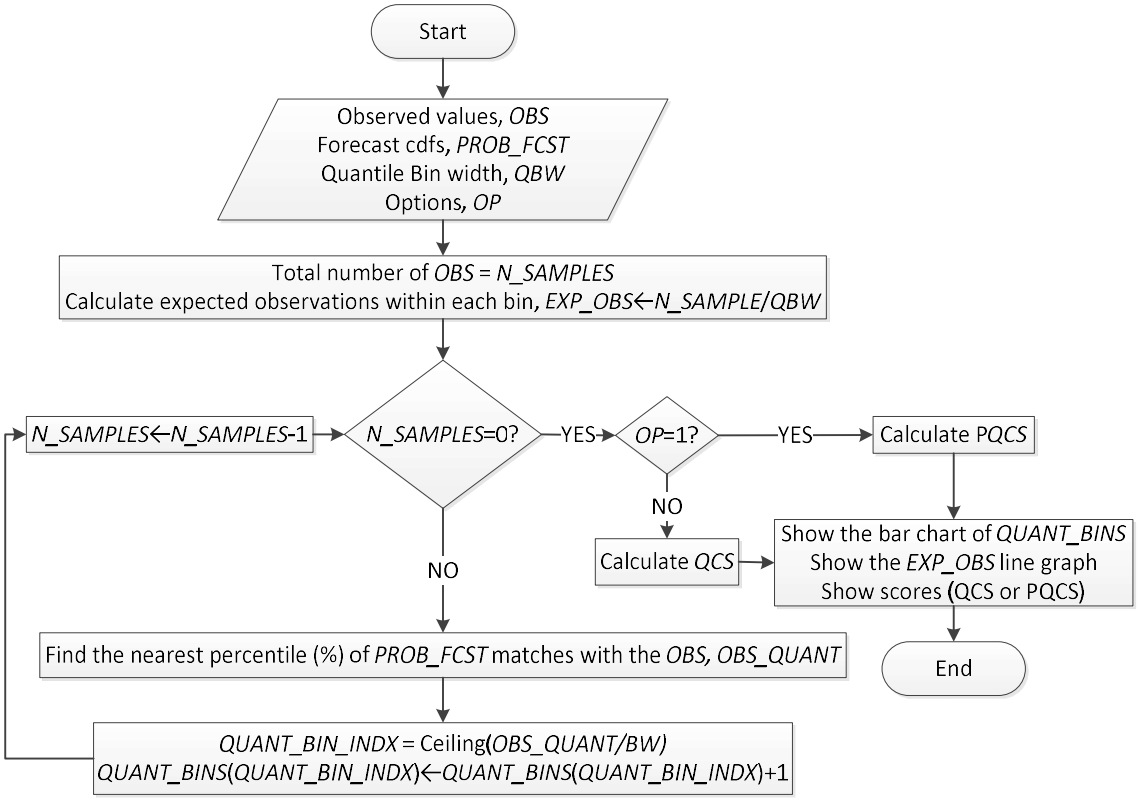

Figure 3.19: Flow chart of the graphical calibration measure evaluation technique.

Figure 3.20 shows a cartoon example of how each observed value is assigned to a quantile bin. The graphical calibration measure finds the nearest percentile of a forecasted CDF from its corresponding observed value. In the cartoon example, the actual

value is close to the 33rd percentile of the CDF. Now, the observed quantile (33rd) is divided by bin width (10) to get the quantile bin index ($\lceil 3.3 \rceil = 4$). Therefore, the observe value is the member of the 4th quantile bin. The same process is repeated for each observed value to fill up all quantile bins.

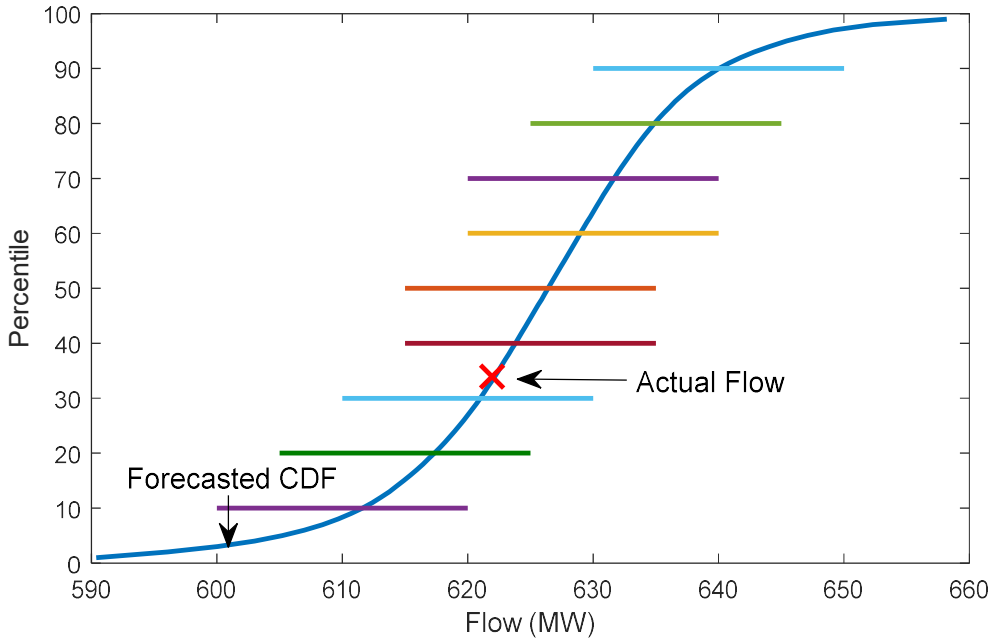

Figure 3.20: Finding nearest percentile (%) of a forecasted CDF from an actual flow.

When the process of assigning each actual value to one of the quantile bins is done, the bar chart of quantile bins looks similar to Figure 3.18 (See also Figure 3.21 for a sample illustration of GCM on a real dataset). It is expected that all ten quantile bins (left side of Figure 3.21) will contain the same number of observations (10% of the total observed values). That means that a bar chart representing frequency of observed values within all (ten) quantile bins should appear as a uniform distribution.

The first and last quantile bins (in the left side of Figure 3.21) may be split into two sub-quantile bins (as shown in the right side of Figure 3.21) to observe carefully the performance of the probabilistic forecast in the tails of the distribution (extreme conditions). All the forecasted CDFs in this dissertation contain 99 values (1% to 99%). That means 1% of the actuals are expected to be less than the 1st percentile value of the forecasted CDF and more than the 99th percentile value of the forecasted CDF, respectively. The red lines in bar charts shown in Figure 3.21 (right side) are the expected frequencies for a particular quantile bin. Hence, if the height of the bar is close to the red line, it is a better probabilistic forecast.

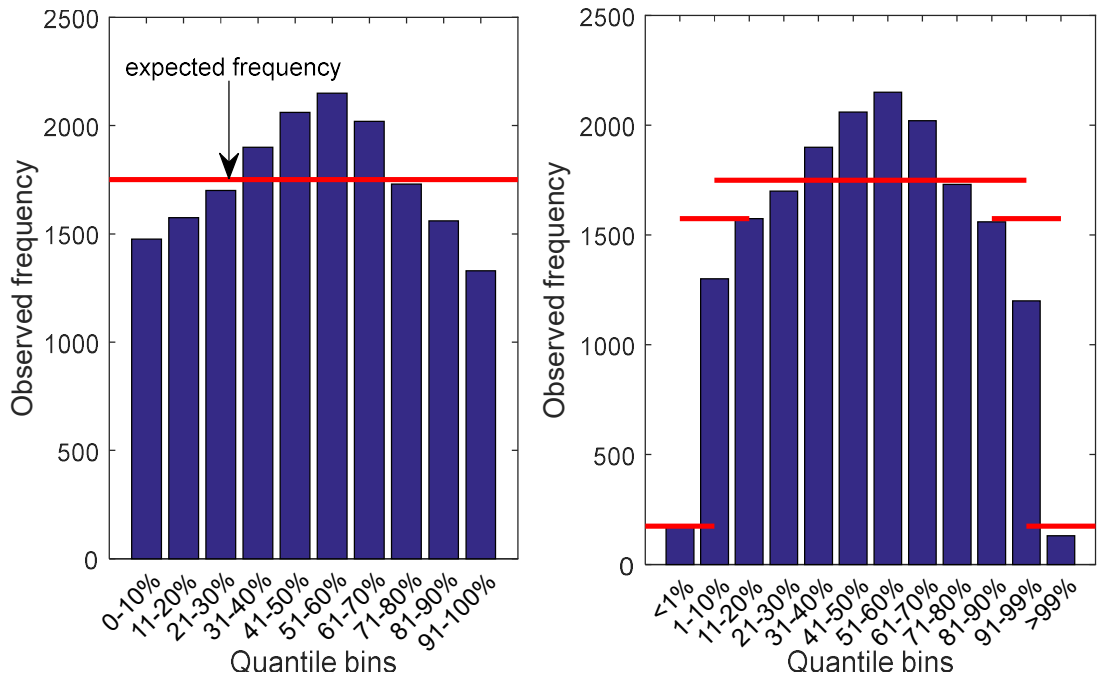

Figure 3.21: Graphical calibration measure (GCM).

Abundant observations in the middle are characteristic of CDFs that are too sharp, and fewer observations in the two tails characterize less sharp probabilistic forecasts (see Figure 3.22). Both too sharp and less sharp probabilistic forecasts are incorrect in terms of forecast uncertainty quantification. Thus, a heavy penalty has been imposed in the primary scoring rule QCS presented in the next subsection (3.5.1) for producing too sharp or less sharp probabilistic forecasts. A dataset-independent version of the QCS called percentage QCS (PQCS) is presented in Subsection 3.5.2.

3.5.1 Quantile Calibration Score (QCS)

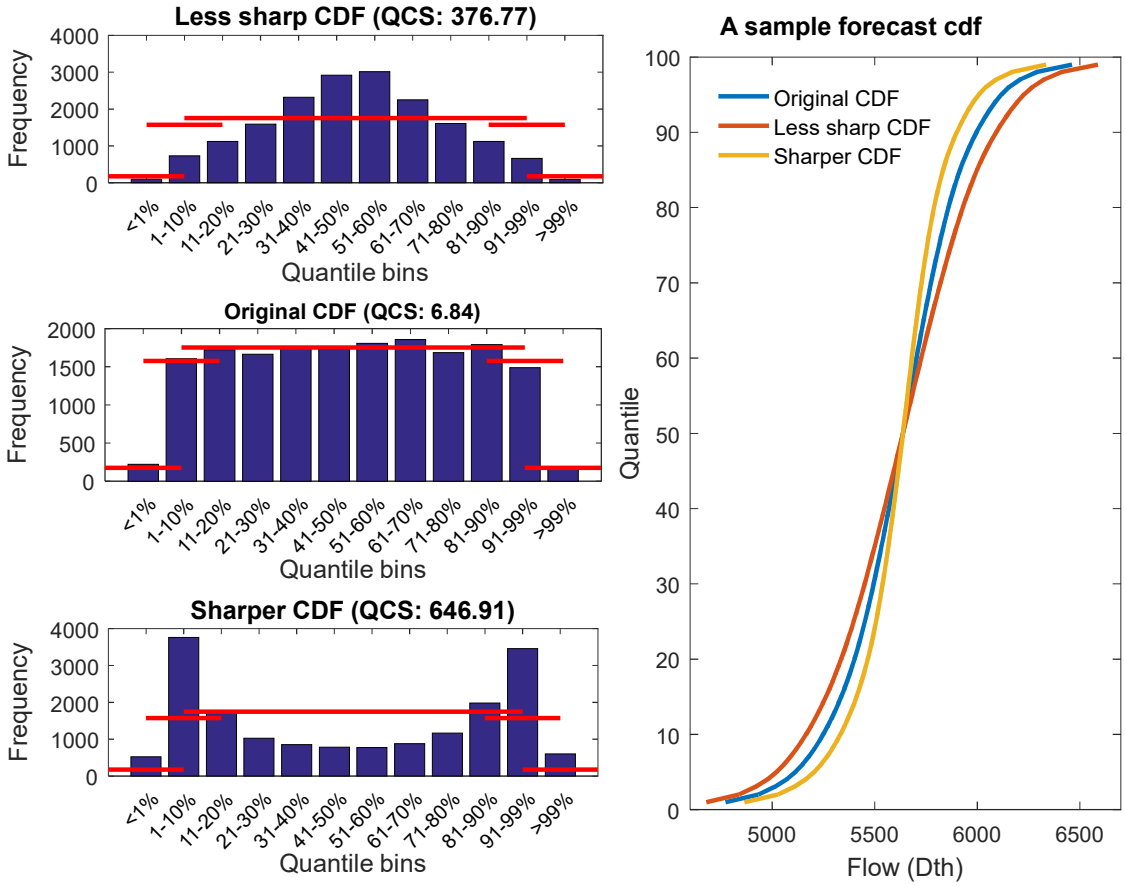

Figure 3.22: Effect of sharper and less sharper CDF on QCS.

If the forecasted CDFs are divided into n quantile bins (in this work, $n = 10$), observed frequencies are calculated by counting the number of observed values within a certain quantile range, and the expected frequency is equal to the total number of observed values divided by n (assuming quantile bins are equal), then the quantile calibration score is

$$QCS = \frac{1}{n}\sum_{bin=1}^{n}\frac{\left(ExpectedFrequency - ObservedFrequency\right)^2}{ExpectedFrequency}. \tag{3.11}$$

Lower QCS are better. The best possible QCS value is zero. A cartoon scenario is created in Figure 3.22 to illustrate the scoring philosophy of the QCS. The original CDF is constructed from the KDE probabilistic forecasting method explained in the Section 3.3. The sharper and less sharper CDFs are created by perturbing the original CDF by $\pm 0.1\%$. The QCS formula was applied to the three forecasted CDFs (original, too sharp, and less sharp) to understand the effect of sharpness and calibration on the QCS. In both cases (too sharp or less sharp), the QCS penalizes a forecast CDF heavily for being too sharp or less sharp and provides a better score to a correct CDF. The next section provides another scoring rule PQCS that is data independent.

3.5.2 Percentage Quantile Calibration Score (PQCS)

When scores are compared between two different size datasets, the QCS is not useful, because it is the measurement of average error. Errors are more when the dataset contains more samples than a smaller size dataset. Thus, a percentile version of the QCS,

called percentage quantile calibration score (PQCS) shown in Equation (3.12), is helpful in this situation.

$$PQCS(\%) = \frac{1}{n}\sum_{bin=1}^{n} \frac{\left|ExpectedFrequency - ObservedFrequency\right|}{ExpectedFrequency} \times 100 . \qquad (3.12)$$

The PQCS is used to do unusual day analysis [135] (see Chapter 4), which is normally 5% of the more difficult days to forecast. Thus, PQCS are used in this dissertation to compare unusual days' scores with all days' scores.

This chapter presented two main contributions of this dissertation: new methods to generate probabilistic forecasts and a new evaluation technique. A competitive point forecasting method, which is used to generate probabilistic forecasts in this work is explained in Section 3.1. Three new probabilistic forecasting methods such as a parametric, a non-parametric, and a semi-parametric method are introduced in Sections 3.2, 3.3, and 3.4, respectively. The second major contribution of this dissertation, a new probabilistic forecast evaluation technique (GCM) with two scoring rules (QCS and PQCS) is offered in Section 3.5. The next chapter contains the application of probabilistic forecasts in the energy industry (natural gas and electricity) and a performance analysis of all proposed methods during normal days as well as unusual days. The GCM and two scoring rules (QCS and PQCS) explained in this chapter are applied to assess probabilistic forecasts. Comparisons of three probabilistic forecasts also are included.

CHAPTER 4

APPLICATION AND ANALYSIS OF PROBABILISTIC FORECASTING METHODS

This chapter applies probabilistic forecasting methods and the evaluation techniques proposed in Chapter 3. Two types of real datasets: electricity and natural gas, collected from two utilities in the U.S. demonstrate the performance of the probabilistic forecasting methods. Evidence of interval forecasts are found in the natural gas industry based on conversations with practitioners [2]. These interval forecasts assume normality. However, based on an extensive literature review, there is no prior evidence of probabilistic forecast uses in the natural gas industry. Thus, showing the application of probabilistic forecasts to solve a real problem in the natural gas industry is the third major contribution of this dissertation. The performance of the point forecasting method used in this dissertation also is analyzed in this chapter. Three variants of each of the three probabilistic forecasting methods are compared. Performance analysis of probabilistic forecasting methods using forecasted weather data as well as actual weather is presented. Unusual days (top 5% difficult days to forecast) analysis for probabilistic forecasts is included. A new evaluation technique, graphical calibration measure (GCM), is used to evaluate probabilistic forecasts. Two new scoring rules, quantile calibration score (QCS) and percentage quantile calibration score (PQCS) introduced in this dissertation (see Section 3.5), are used to assess probabilistic forecasts along with two well-known scoring rules, pinball score and continuous ranked probability score (CRPS) (see Section 2.3).

4.1 Point Forecast Result Analysis

This section analyzes the performance of the point forecasting method, multiple linear regression 3 (MLR3) (see Section 3.1), used in this dissertation to generate probabilistic forecasts. Two different datasets (electricity and natural gas) used for this research are collected from an electricity distribution company and a natural gas distribution company located in the U.S. Weather data used for generating forecasts are also collected from those two utility companies. All electricity and natural gas demand data are scaled to preserve confidentiality.

For both electricity and natural gas datasets, 12 years of hourly data are used to create seven subsets as shown in Table 4.1. The first subset contains five years of training data and the following one year of testing data. The second subset adds one year to the previous five years training dataset for creating a new training dataset. Similarly, the third to seventh training datasets are built upon the previous training dataset by adding one additional year of data. The length of the testing dataset is one year for all seven datasets. This strategy is more realistic than using fixed length training datasets because the length of the available dataset increases with time in practice, and including more data in the training dataset is helpful to improve forecasting performance. In the GasDay^{TM} lab, the frequency of receiving new data is around 24 hours. However, creating a new forecasting model using additional data is a costly operation. Thus, the GasDay^{TM} lab rebuilds their forecasting models once a year. A similar approach is taken in this work to ensure the performance analysis of forecasts reflects practice.

Data used in this work were detrended using Brown et al.'s detrending algorithm [207]. Brown et. al. showed that detrending improves point forecasts. In this work, evidence of improvement in probabilistic forecasts was found by using a detrended dataset. Thus, detrended datasets are used for both probabilistic electricity and natural gas forecasting.

Table 4.1: Training and testing subsets for the MLR3 method.

	Year							
	1-5	6	7	8	9	10	11	12
Subset 1	Training	Testing						
Subset 2	Training		Testing					
Subset 3	Training			Testing				
Subset 4	Training				Testing			
Subset 5	Training					Testing		
Subset 6	Training						Testing	
Subset 7	Training							Testing

Two popular point forecasting evaluation techniques, mean absolute percentage error (MAPE) and root mean squared error (RMSE) [173], are used in this work to analyze the performance of both electricity and natural gas point demand forecasts. MAPE is helpful to compare the forecasting performance of multiple datasets (or utilities) and forecasting models. On the other hand, RMSE is useful to know the average forecasting error. If Y_t are the observed (actual) values, $\hat{Y}_t$ are the forecasted values at time *t*, and *n* is the number of instances, then the MAPE and RMSE are given by (4.1) and (4.2), respectively.

$$MAPE = \frac{100}{n}\sum_{t=1}^{n}\left|\frac{Y_t - \hat{Y}_t}{Y_t}\right|. \tag{4.1}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(Y_t - \hat{Y}_t\right)^2}. \tag{4.2}$$

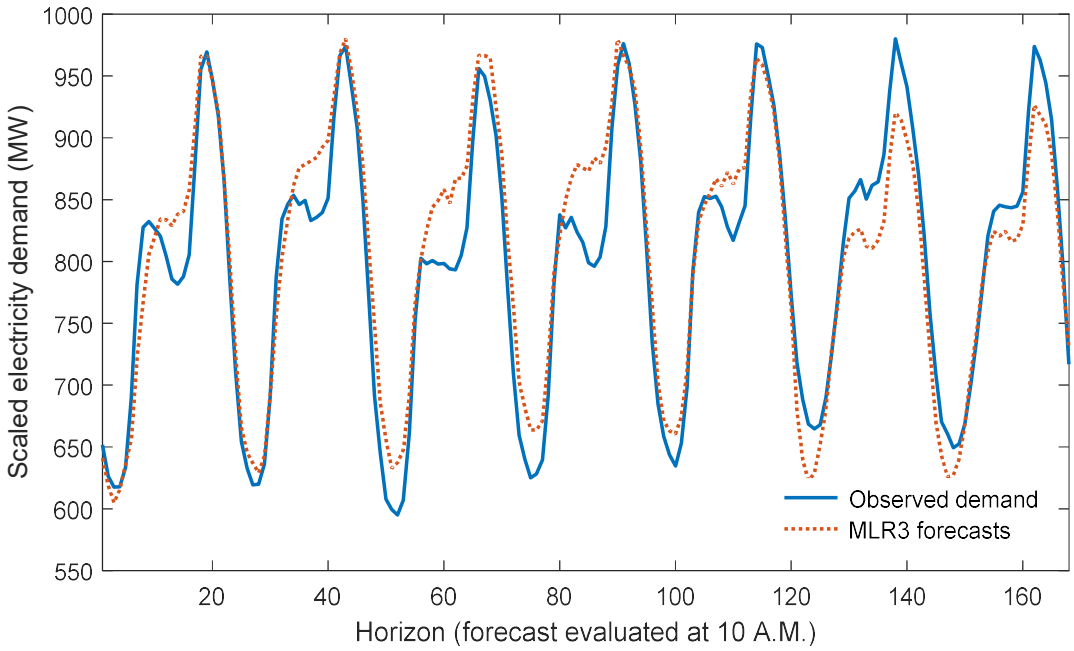

Figure 4.1: A sample electricity demand forecast for a one to 168 hour time horizons.

Figure 4.1 shows a typical pattern of a week long (168 hours) electricity demand actual and forecast loads (using the MLR3 method). In this work, seven subsets (see Table 4.1) are used to train seven separate years of point forecasts. Residuals calculated from those seven years of point forecasts are used to generate probabilistic forecasts.

The Kolmogorov-Smirnov (K-S) test checks whether the forecasting results found from different subsets are significantly different. The hypothesis (H_0) and the alternative hypothesis (H_a) for this test are given below.

H_0: “the forecasts obtained from different subsets in Table 4.1 come from the same distribution.”

H_a: “the forecasts obtained from different subsets in Table 4.1 come from different distributions.”

The statistical test result does not reject the null hypothesis with 5% significance level for all 21 pairs of subsets (two subsets are selected at a time from seven subsets). Similar results are found for both electricity and natural gas datasets. Thus, it is conclusive that forecasts calculated from different subsets (Table 4.1) come from the same distribution. Hence, the MAPE calculated from different subsets can be used to train or test probabilistic forecasting models discussed in Sections 3.2, 3.3, and 3.4.

For electric load forecasts, the seven-year average MAPE of the MLR3 method (Table 4.2) for 1 hour, 1 day, and 1 week time horizons are 1.3%, 4.8%, and 6.7%, respectively. This is comparable to the other two linear regression methods (MLR1 and MLR2) discussed in Section 3.1. The performance analysis of MLR1 and MLR2 methods are not included in this dissertation because the main focus of this work is to provide useful probabilistic forecasting methods and evaluation techniques. The average over the seven subsets from one to 168-hour horizon MAPE and RMSE of the point forecasting model MLR3 are presented in Figure 4.2. After the first 48 hours, the forecast error plateaus.

Table 4.2: Yearly MAPE and RMSE of electricity demand forecasts using the MLR3.

	1-hour horizon		24-hour horizon		168-hour horizon	
	MAPE (%)	RMSE (MW)	MAPE (%)	RMSE (MW)	MAPE (%)	RMSE (MW)
2009-10 (Subset 1)	1.32	10.59	4.79	39.84	6.92	58.79
2010-11 (Subset 2)	1.32	11.80	4.71	42.45	6.22	56.28
2011-12 (Subset 3)	1.32	10.89	5.07	42.86	6.94	58.07
2012-13 (Subset 4)	1.33	11.41	4.59	38.74	6.23	55.16
2013-14 (Subset 5)	1.25	10.72	4.59	39.28	6.73	56.67
2014-15 (Subset 6)	1.16	9.58	4.87	41.14	6.67	54.10
2015-16 (Subset 7)	1.13	9.00	4.76	39.19	7.06	60.63
Average	**1.26**	**10.57**	**4.77**	**40.50**	**6.68**	**57.10**

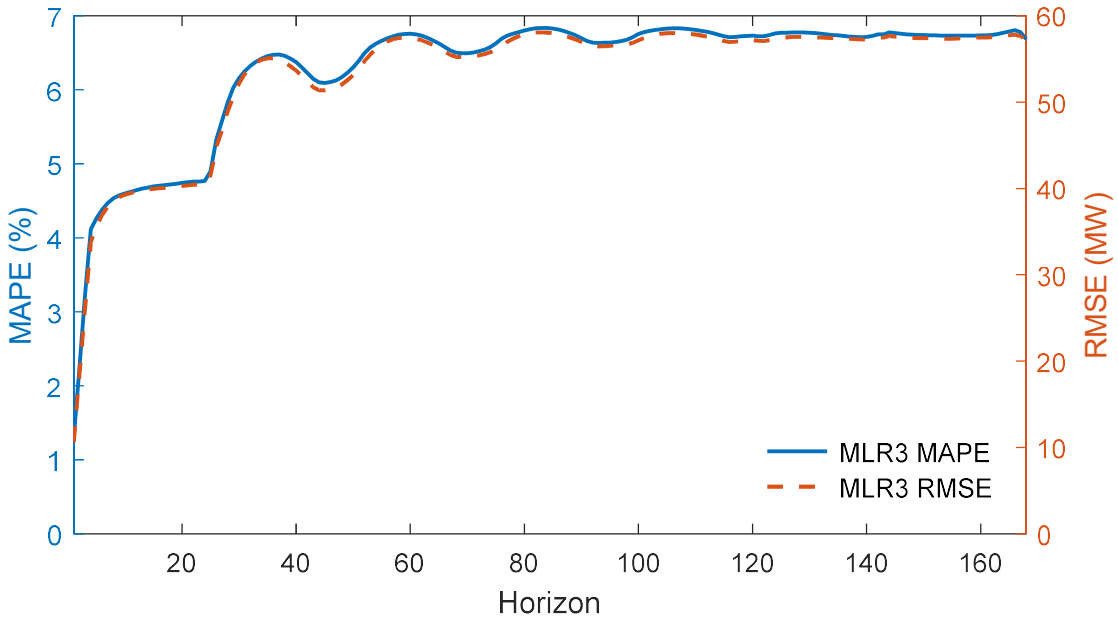

Figure 4.2: Seven year average, one to 168 hour horizon MAPE and RMSE calculated from the detrended electricity demand dataset.

The same MLR3 method used for electricity load forecasting can be used to forecast natural gas demand since both natural gas and electricity demand depend on historical weather, seasonality, and historical load. Figure 4.3 shows a typical pattern of a week-long hourly natural gas forecasts (using the MLR3 method) and observed natural gas flow. The performance of the one week ahead natural gas forecast is not as accurate as the electricity demand forecasts because of potential outliers and missing input variables in the MLR3 model specific to natural gas demand. In addition, natural gas demand is more variable than electricity demand based on two case studies presented in this dissertation. However, improving point forecasts is not the main concern of this work. When point forecasts are poor, then probabilistic forecasts may be more useful. Thus, the same MLR3 method designed for electricity forecasting is used to forecast natural gas.

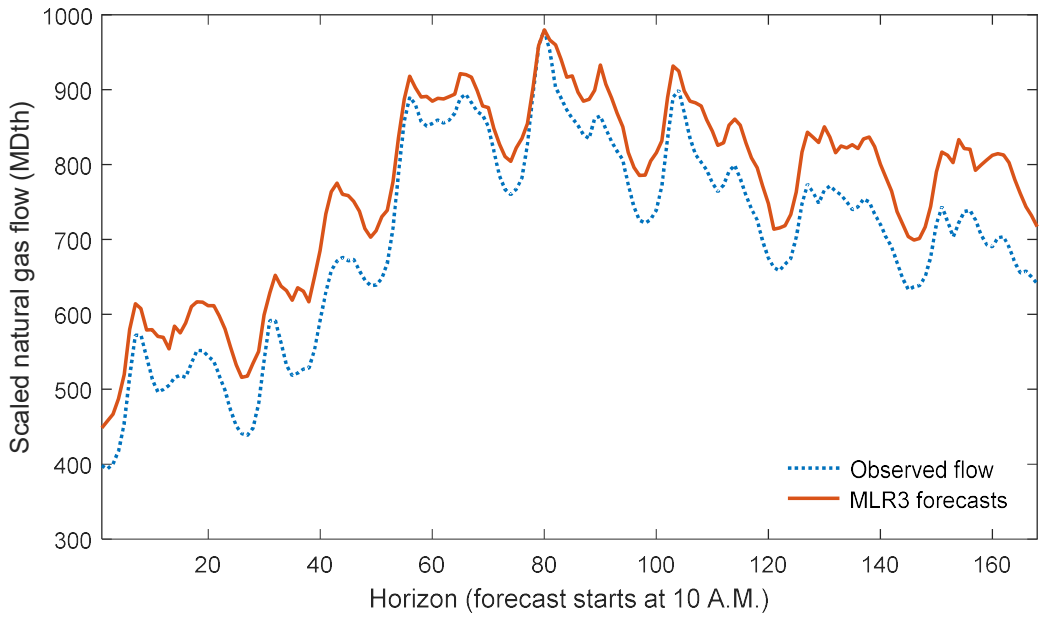

Figure 4.3: A sample week-long hourly natural gas flow point forecasts.

Natural gas data processing is done in an analogous manner to that of the electricity data processing as shown in Table 4.1. Seven separate years of natural gas forecasts are generated from seven subsets. Table 4.3 shows the seven yearly MAPE and RMSE of natural gas forecasts for 1 hour, 1 day, and 1-week time horizons using the MLR3 method. The average seven-year average MAPE for 1 hour, 1 day, and 1-week horizons are 2.7%, 16.9%, and 24.5%, respectively. Figure 4.4 shows a week-long hourly MAPE and RMSE of natural gas forecasts. The point forecasts generated in this section will be used to generate probabilistic forecasts in the next section.

Table 4.3: Yearly MAPE and RMSE of natural gas flow forecasts using the MLR3.

	1-hour horizon		24-hour horizon		168-hour horizon	
	MAPE (%)	RMSE (MW)	MAPE (%)	RMSE (MW)	MAPE (%)	RMSE (MW)
2009-10 (Subset 1)	2.64	577.98	15.20	2079.95	22.84	2860.30
2010-11 (Subset 2)	2.95	641.59	17.00	2499.81	24.42	3221.53
2011-12 (Subset 3)	2.89	529.96	18.40	2338.60	28.30	3369.06
2012-13 (Subset 4)	2.61	531.48	16.33	2355.30	22.65	3066.01
2013-14 (Subset 5)	2.53	564.40	17.19	2584.00	24.94	3479.79
2014-15 (Subset 6)	2.44	509.86	16.45	2515.53	23.16	3412.81
2015-16 (Subset 7)	2.86	580.88	17.40	2459.41	25.12	3225.90
Average	**2.70**	**562.31**	**16.85**	**2404.66**	**24.49**	**3233.63**

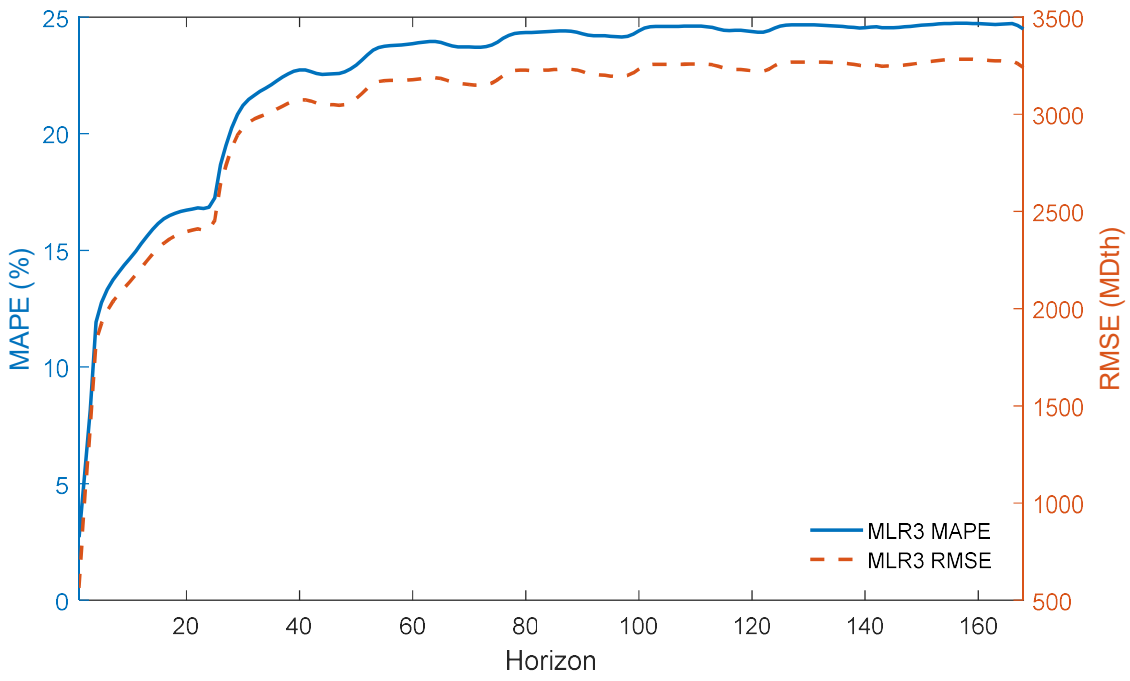

Figure 4.4: Seven years average, one to 168-hour horizon MAPE and RMSE calculated from the detrended natural gas flow dataset.

The MLR3 point forecasting method is used to generate probabilistic forecasts in this work for both electricity and natural gas datasets. The next section uses the MLR3 point forecasts generated in this section as an input to produce probabilistic forecasts. The following sections provide weather forecast based probabilistic forecasts and performance analysis of probabilistic forecasts during unusual days (top 5% difficult days to forecast).

4.2 Probabilistic Forecasting Results

This section presents probabilistic forecasts generated from three different methods (NDEPF, KDEPF, and JDTPF) discussed in Sections 3.2, 3.3, and 3.4, respectively. Each of the three methods has three variants (see Equation (3.3) and (3.4), Section 3.2). Seven years of electricity demand / natural gas flow data are used for training and testing purposes. Historical point forecasting errors required to train probabilistic forecasting methods are collected from the MLR3 point forecasts (Section 4.1). Historical observed weather data is used in this section to generate probabilistic forecasts; the next section shows probabilistic forecasts using forecasted weather data. Table 4.4 shows the data processing summary for creating training and testing datasets for probabilistic forecasts. Two subsets are created to train probabilistic forecasts. The first subset contains five years of hourly MLR3 point forecast errors and historical temperatures for training probabilistic forecasting models. The second subset contains six years of training data. Both subsets have one year of testing data. The evaluation techniques explained in Section 3.5 are applied on two years of hourly probabilistic forecasts (total 17,520 probabilistic forecasts, excluding leap hours) to assess the performance of the new probabilistic forecasting methods. Examples of 24-hour horizon probabilistic forecasts are illustrated in this section.

Table 4.4: Data processing for probabilistic forecasts.

	Year

	1-5	6	7	8	9	10	11	12
Subset 1		Training					Testing	
Subset 2		Training						Testing

Figure 4.5 shows a sample day-long hourly electricity demand forecast using the Johnson data transformation probabilistic forecast (JDTPF) method. The kernel density estimator probabilistic forecast (KDEPF) method produces similar probabilistic forecasts, and the normal distribution estimator probabilistic forecast (NDEPF) method typically generates sharper probabilistic forecasts (Figure 2.2) than JDTPF and KDEPF methods (not shown in figures). The probabilistic forecast generated for each horizon is a cumulative distribution function (CDF), where the 50^{th} quantile can be denoted as the point forecast (see inset of Figure 4.5). In this work, 99 quantiles are displayed using 50 distinct colors. In Figure 4.5, the observed demand is more than the point forecast (50^{th} quantile) until 10 P.M. and less than the point forecast afterwards. In the long run, actuals are expected to be more than the 50^{th} quantile half of the time and less than the 50^{th} quantile half of the time. The electricity demand is less than the 10^{th} quantile (orange colored line) for last few hours (7-9 A.M.). If 7-9 A.M. was repeated many times, then actual demands are expected to be greater than the 10^{th} quantile 90% of the time and less than the 10^{th} quantile 10% of the time.

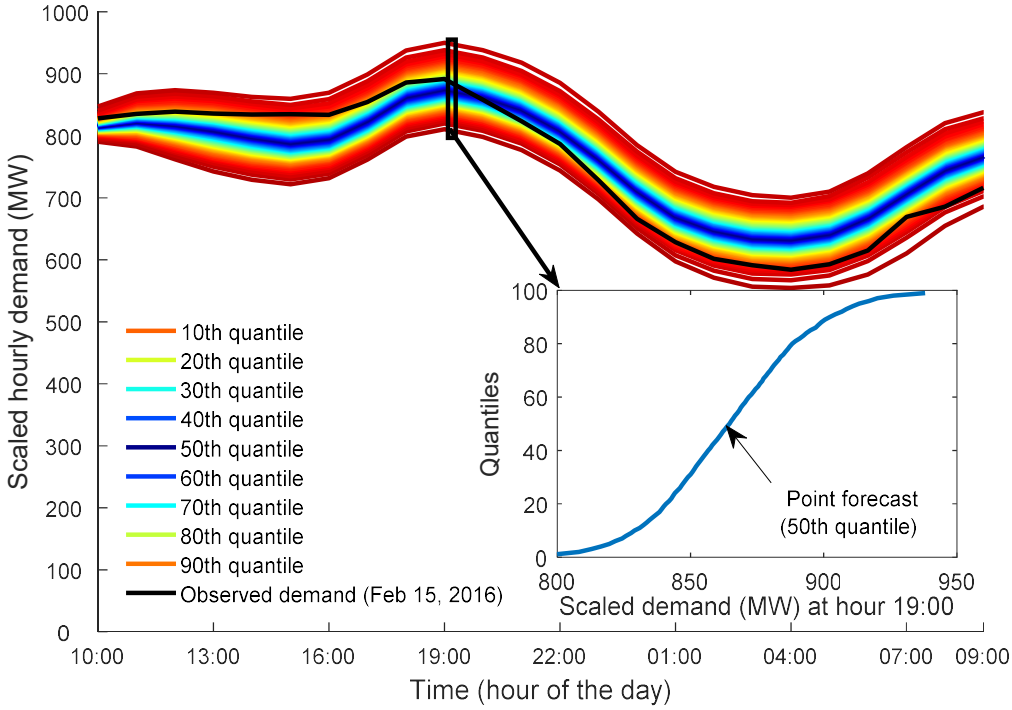

Figure 4.5: A sample day-long hourly electricity demand probabilistic forecasts.

Probabilistic forecasts generated from the natural gas dataset contain more features to analyze than the electricity dataset generated probabilistic forecasts because of greater variability in the natural gas dataset. In the natural gas dataset, the highest flow is 49 times the lowest flow, whereas the highest demand in the electricity dataset is only four times the lowest demand. Thus, most of the figures shown in this section to illustrate probabilistic forecasts are natural gas probabilistic forecasts.

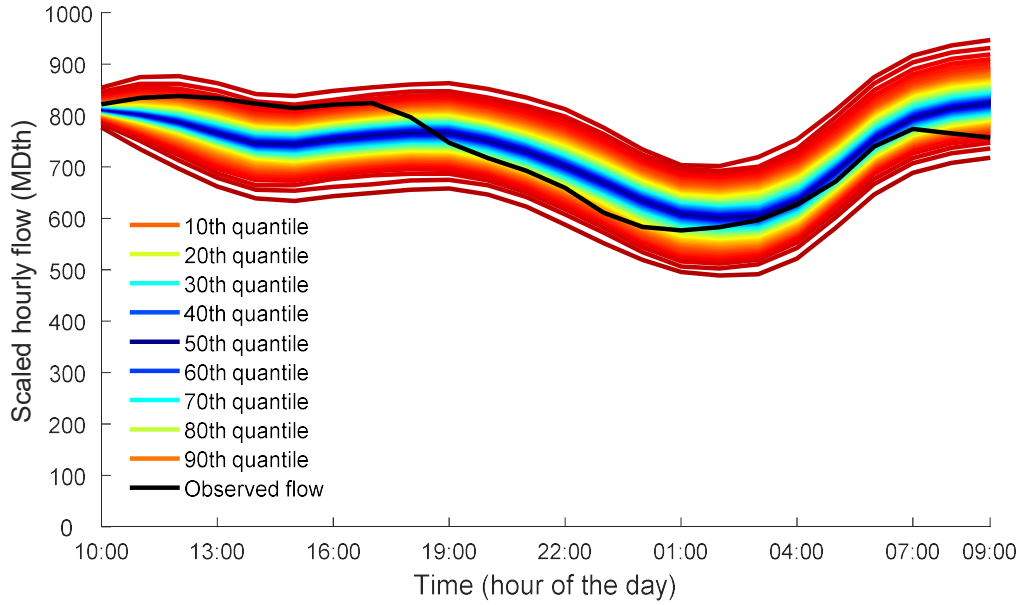

Figure 4.6: A sample day-long hourly natural gas probabilistic forecast, where actual flow swings between the 3rd quantile and the 98th quantile (Date: Dec 28, 2015).

Figure 4.6 shows a sample natural gas probabilistic forecast, where actual flow swings between the 3rd quantile (at 9 A.M.) and 98th quantile (at 2 P.M.). The point forecasts are accurate (actual flow is close to the 50th quantile) between 4 P.M. and 7 P.M. only and inaccurate most of the time horizon due to uncertainty. A probabilistic forecast is a helpful tool to quantify the forecast uncertainty. The probabilistic forecast in Figure 4.6 is sharp (see Figure 2.2) compared to the one in Figure 4.7, which indicates that every probabilistic forecast has a different resolution (forecasted CDF changes with time). Resolution is considered as one of the main criteria of good probabilistic forecasts [37]. Larger differences between two quantiles (Figure 4.7) mean more uncertainty compared to smaller differences between two quantiles (Figure 4.6). It is expected that the differences between two quantiles will grow with increasing time horizon because the amount of uncertainty typically increases with the length of the time horizon. The

uncertainty also comes from weather forecasts and human behavior in addition to the time horizon. Thus, it is possible to have sharper CDFs with the increase of the time horizon. A fixed difference between any two quantiles over time (no resolution) is considered as a poor probabilistic forecast, since it does not capture the variability of the uncertainty [37].

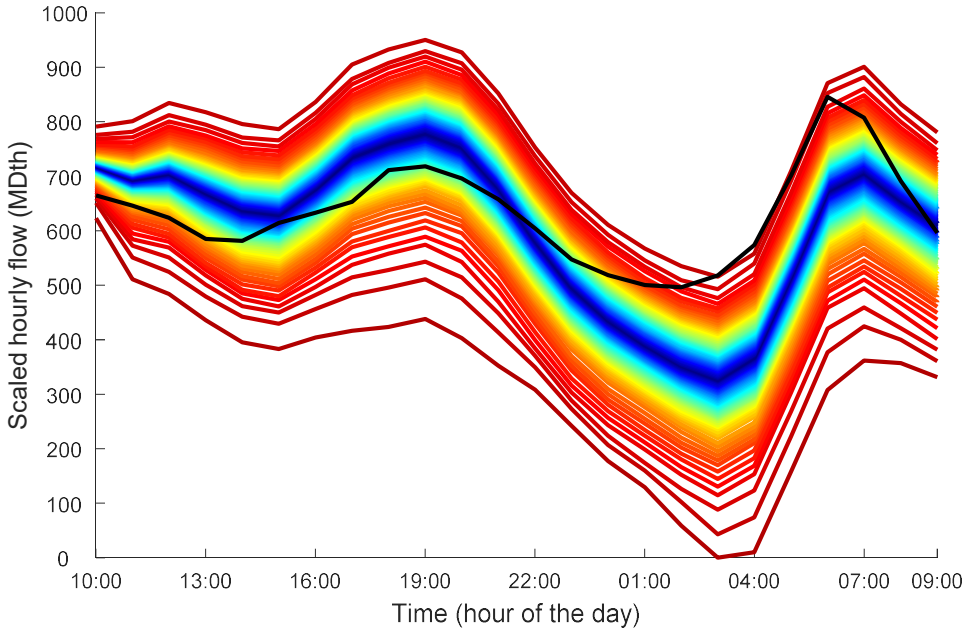

Figure 4.7: A sample day-long hourly probabilistic forecast, where forecasted CDFs are less sharp than usual indicates more uncertainty (Date: Apr 18, 2016).

In Figure 4.8, the observed flow touches the 1st quantile at 5 P.M. It is expected that 1% of the time, the actual flow will be below that point, and 99% of the time the actual flow will be above that point. The point forecasts are made during the heating season (Jan 30 at 10 A.M., 2016), when the largest amount of natural gas is used for heating purposes. A large temperature swing (between 25^{0}F and 50^{0}F) within the last 24

hours caused poor point forecasts. However, the probabilistic forecast nicely captures the uncertainties involved with the point forecast.

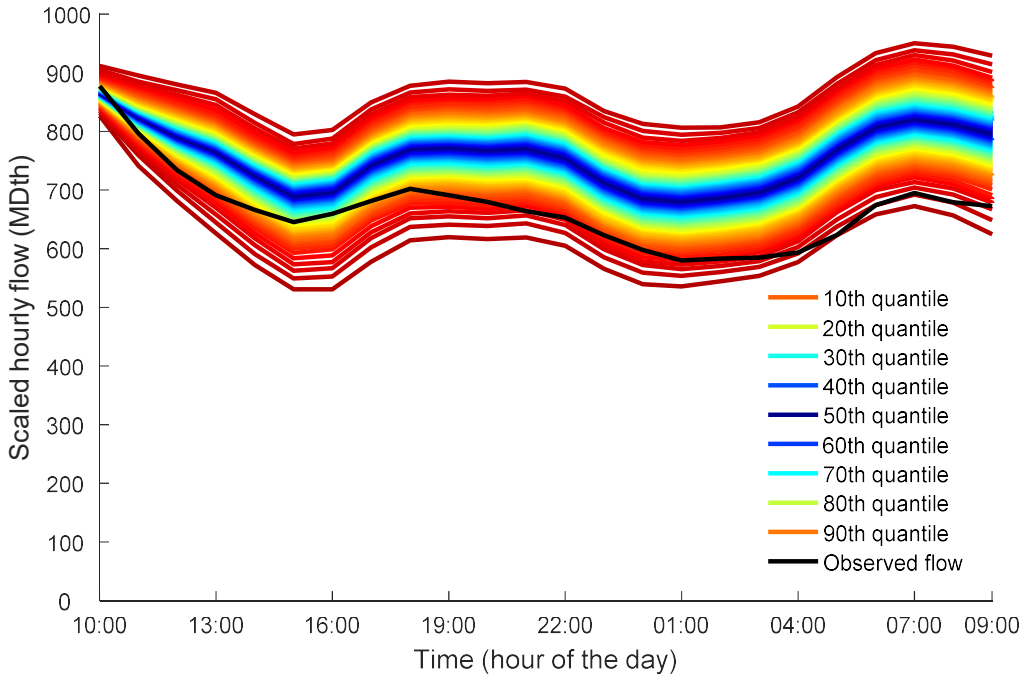

Figure 4.8: A sample day-long hourly natural gas probabilistic forecast which touches the 1st quantile at 5 P.M., Jan 30, 2016.

In Figure 4.9, the actual flow is above the 99th quantile for two consecutive hours (7 A.M. – 9 A.M.). A sudden sharp drop in temperature from 62^0F to 29^0F within a very brief period triggered this poor point forecast. In this case, the probabilistic forecast seems poor because the actual flow is outside the maximum level of the forecasted CDF (99th quantile), but it is not. It is expected that 1% of the time, the actual flow will be more than the 99th quantile. Similarly, the actual flow is expected to be less than the 1st quantile 1% of the time (Figure 4.10). From a single probabilistic forecast, it is not

possible to conclude whether it is a bad or good probabilistic forecast. A series of probabilistic forecasts is required for evaluation. The next four subsections (4.2.1, 4.2.2, 4.2.3, and 4.2.4) present probabilistic forecast evaluation results.

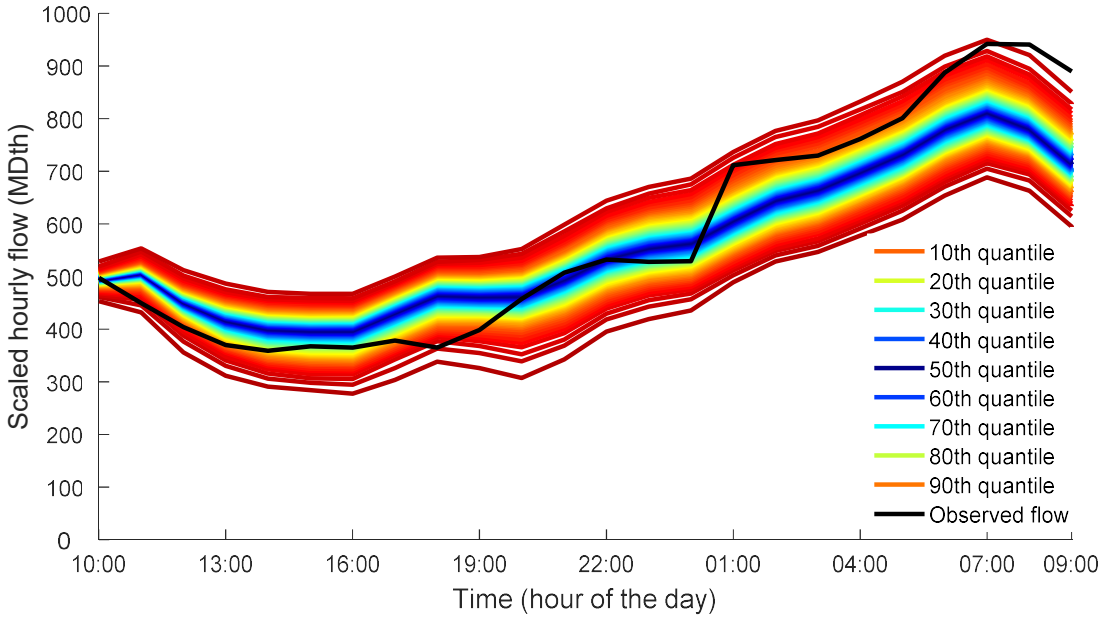

Figure 4.9: A sample day-long hourly natural gas probabilistic forecast where the actual flow is outside the 99th quantile for two consecutive hours (7-9 A.M., Feb 28, 2016).

Figure 4.11 shows another incident where the actual flow crosses the 99th quantile (1 P.M. - 4 P.M.). Higher quantiles (51st to 99th quantiles) usually are more important to gas controllers than lower quantiles (1st to 49th quantiles), because gas controllers are responsible for keeping the actual flow below the contractual maximum flow limit (Figure 1.1), which is set before 10 A.M. every day [2]. Lower quantiles also are useful to set the daily cumulative minimum flow limit (Figure 1.2). While setting those important limits, one should keep in mind that the 1st quantile and the 99th quantile are

not the minimum and maximum actual flow boundaries, respectively. 2% of the time, the actual flow is expected to be outside these boundaries.

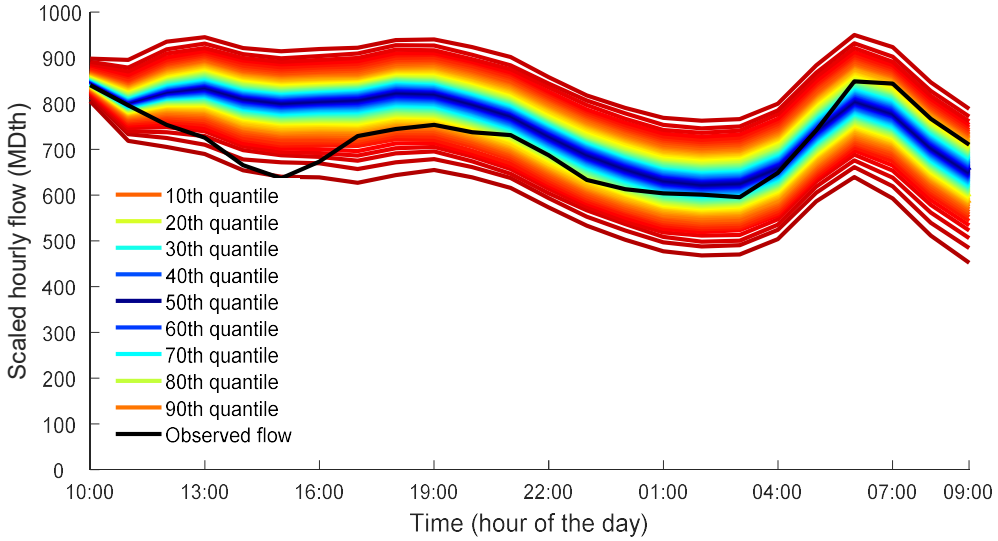

Figure 4.10: A sample day-long hourly natural gas probabilistic forecast where the actual flow touches the 1st quantile at 3 P.M., Apr 4, 2016.

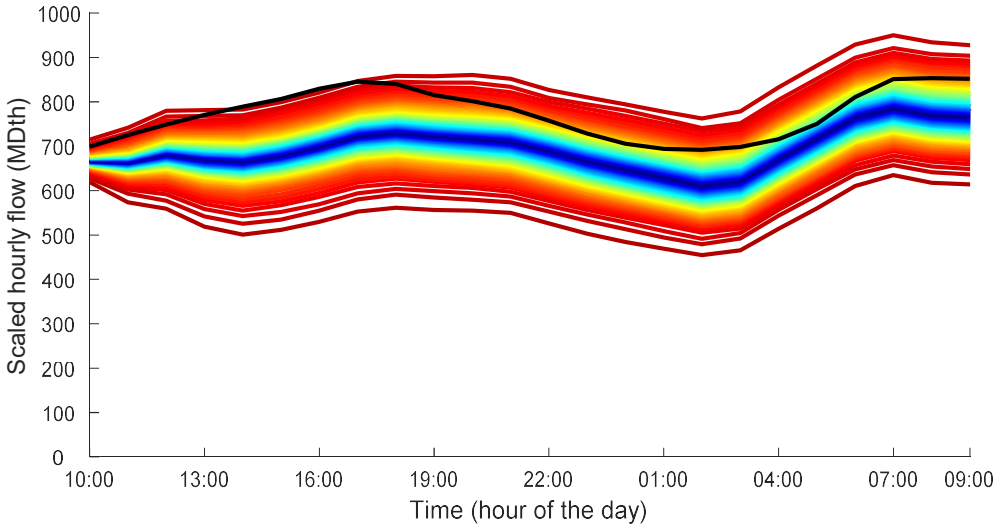

Figure 4.11: A sample day-long hourly probabilistic forecast, where the actual flow is very close to the 99th quantile for three consecutive hours (2-5 P.M., Nov 27, 2015).

The new probabilistic forecast scoring rules, the quantile calibration score (QCS) and the percentage quantile calibration score (PQCS) require a series of probabilistic forecasts to calculate a score. Thus, 137 groups are created from 17,520 probabilistic forecasts, generated from two testing subsets (see Table 4.4). Each of the groups contains 1200 probabilistic forecasts, and two consecutive groups overlap 90% of their probabilistic forecasts (see Figure 4.12). The parameters of creating groups such as the number of probabilistic forecasts in a group, and the percentage of overlapping between two consecutive groups are changed to try different ways of grouping. Scores obtained from grouping probabilistic forecasts in different ways are similar. The QCS and the PQCS provided in the next subsection are obtained by averaging QCS and PQCS calculated from 137 groups.

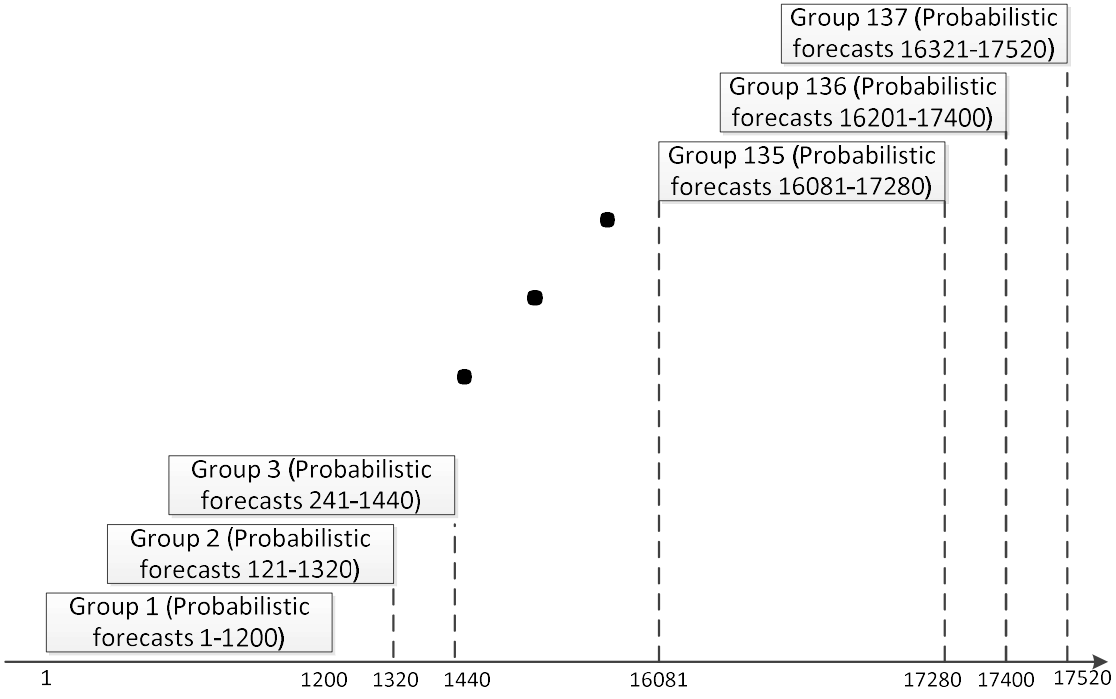

Figure 4.12: Grouping for probabilistic forecast evaluation using QCS and PQCS.

The rest of this section presents performance analysis of three probabilistic forecasting methods (NDEPF, KDEPF, and JDTPF) in three different subsections (Sections 4.2.1, 4.2.2, and 4.2.3). Finally, the last subsection (Section 4.2.4) compares the performance of the three probabilistic forecasting methods.

4.2.1 Performance Analysis of the Benchmark Method, NDEPF

This section presents the performance analysis of the benchmark method, the normal distribution estimator probabilistic forecast (NDEPF). Three variants (Section 3.2) are used to generate probabilistic forecasts using the NDEPF. One to 168 hours (1 week) horizon probabilistic forecasts have been generated to compare the performance of three variants using four probabilistic forecasting scoring rules (Figures 4.13 and 4.14). Some sample 24 hour probabilistic forecasts are shown in the previous section. Table 4.5 (electricity dataset) and Table 4.6 (natural gas dataset) show the performance of the NDEPF based on four scoring rules, pinball score, continuous ranked probabilistic score (CRPS), quantile calibration score (QCS), and percentage quantile calibration score (PQCS) for one hour, one day, and one-week horizons. A lower score is better, and the best possible score is zero. The best score among the three variants is highlighted in a bold font. Figures 4.13 and 4.14 show one to 168-hour horizon scores calculated from the electricity dataset and natural gas dataset, respectively.

Table 4.5: Score comparison of three variants of the NDEPF using the electricity dataset.

		Benchmark method, NDEPF (electricity data)		
		Temperature	Temperature difference	Weekly temperature difference
Pinball score	1 hour	**2.53**	2.58	2.57
	24-hour	**9.56**	10.80	10.85
	168-hour	**12.62**	15.65	15.45
CRPS	1 hour	**5.00**	5.10	5.09
	24-hour	**18.93**	21.38	21.48
	168-hour	**25.00**	31.00	30.60
QCS	1 hour	**13.81**	16.70	16.35
	24-hour	**11.06**	19.40	18.87
	168-hour	**21.60**	30.18	29.99
PQCS (%)	1 hour	**29.16**	32.48	32.09
	24-hour	**23.49**	32.75	31.17
	168-hour	**32.45**	41.38	40.70

The temperature variant performed better than other two variants (daily temperature difference and weekly average temperature difference) for the electricity dataset (Figure 4.13). However, weekly temperature difference is a better variant for the natural gas dataset (Figure 4.14) considering QCS and PQCS.

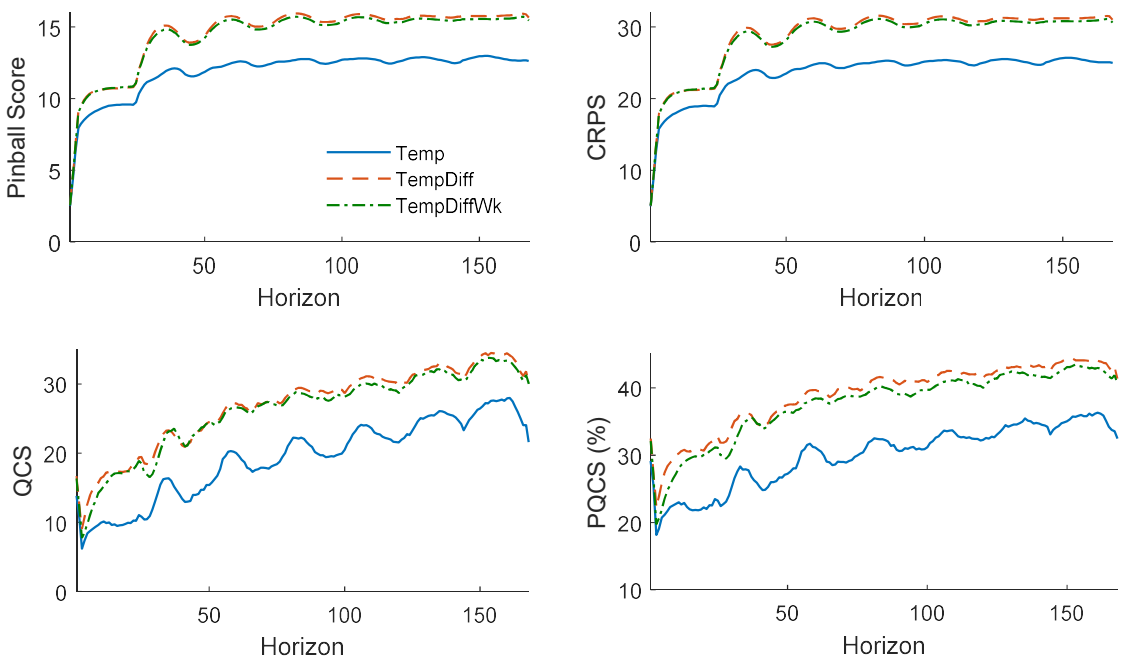

Figure 4.13: Two years average week-long hourly pinball scores, CRPS, QCS, and PQCS for three variants, NDEPF calculated from the electricity dataset.

The trend in Figure 4.13 shows that the probabilistic forecasting scores increase with the time horizon. That means, quantifying probabilistic forecasts is difficult for greater time horizons. The pinball score and the CRPS of the natural gas dataset (Table 4.6) are poor compared to the electricity dataset (Table 4.5), which indicates that these two scoring rules produce better probabilistic forecasting scores when point forecasts are superior. Pinball score, CRPS, and QCS are dataset dependent. Only PQCS is dataset independent, which is helpful to compare the performance of probabilistic forecasts from two different datasets or methods. According to the PQCS, probabilistic forecasts made from the natural gas dataset are better than those made for the electricity dataset.

Table 4.6: Score comparison of three variants of the NDEPF from the natural gas dataset.

		Benchmark method, NDEPF (natural gas data)		
		Temperature	Temperature difference	Weekly temperature difference
Pinball score	1 hour	**138.50**	143.04	142.24
	24-hour	**606.47**	674.41	680.99
	168-hour	**751.24**	949.62	941.03
CRPS	1 hour	**274.06**	282.97	283.39
	24-hour	**1193.82**	1312.25	1337.68
	168-hour	**1491.22**	1948.82	1926.89
QCS	1 hour	**17.21**	27.57	27.60
	24-hour	11.83	**5.88**	6.68
	168-hour	20.75	**14.37**	14.70
PQCS	1 hour	**32.24**	37.69	36.82
	24-hour	25.23	**16.43**	18.19
	168-hour	32.00	26.75	**26.35**

The sharp drop of QCS and PQCS for the first few hours (Figure 4.14) can be explained from the point forecasting results. Point forecasts for the first three-hour horizon are far better than the fourth hour horizon and onwards, because the MLR3 method uses the first three hours as lag terms (Equation (3.1)). Hence, most of the time, it is expected to get sharp probabilistic forecasts (actuals are very close to the 50^{th} quantile, between the 40^{th} and the 60^{th} quantiles), which is not good according to QCS and PQCS. For a good probabilistic forecast, it is expected that the actual flow will be between the 40^{th} and the 60^{th} quantiles 20% of the time and outside this region 80% of the time.

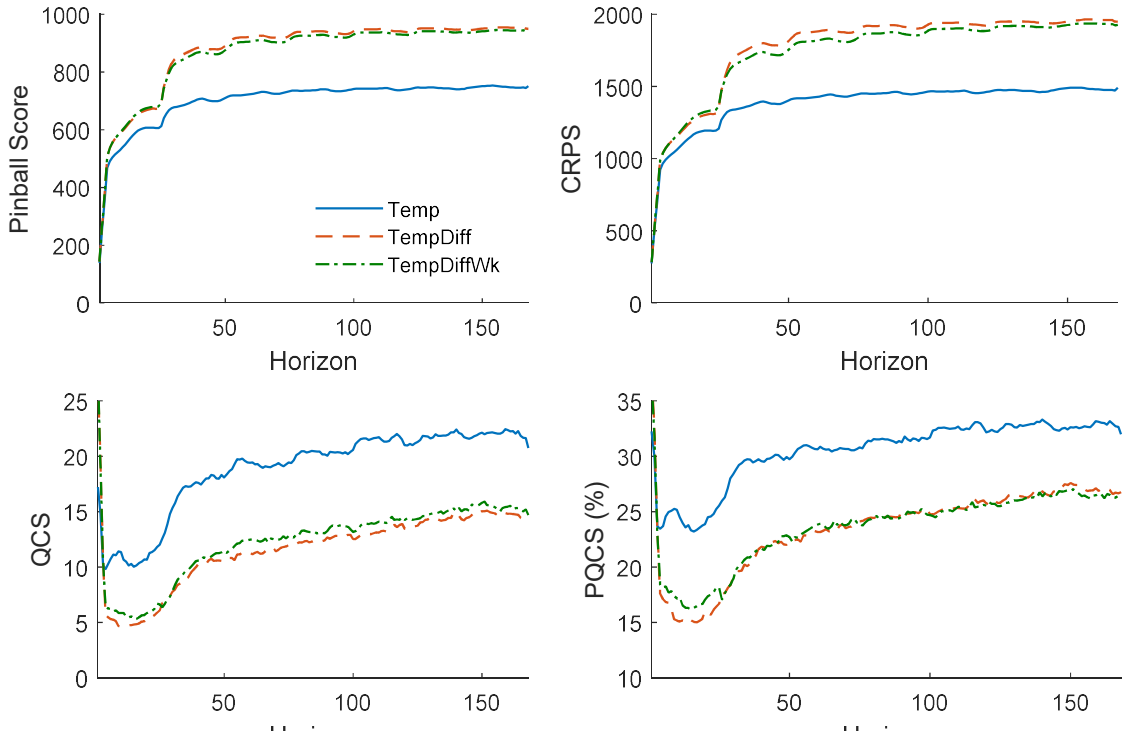

Figure 4.14: Two years average week-long hourly pinball scores, CRPS, QCS, and PQCS for three variants, NDEPF calculated from the natural gas dataset.

4.2.2 Performance Analysis of the KDEPF Method

This section shows the performance analysis of the kernel density estimator probabilistic forecast (KDEPF) method. Table 4.7 and 4.8 show one hour, one day, and one-week horizon scores calculated from the electricity dataset and natural gas dataset, respectively. Three variants are used to generate KDEPF. The temperature variant provides better performance compared to other variants, especially for the electricity dataset (Figure 4.15). However, daily temperature difference shows better performance for the natural gas dataset using the QCS and PQCS (Figure 4.16).

Table 4.7: Score comparison of three variants of the KDEPF using the electricity dataset.

		KDEPF method (electricity data)		
		Temperature	Temperature difference	Weekly temperature difference
Pinball score	1 hour	**2.49**	2.53	2.53
	24-hour	**9.54**	10.62	10.75
	168-hour	**12.62**	15.50	15.30
CRPS	1 hour	**4.93**	5.01	5.01
	24-hour	**18.90**	21.04	21.29
	168-hour	**25.00**	30.71	30.32
QCS	1 hour	**4.91**	5.06	5.51
	24-hour	**8.72**	13.56	12.94
	168-hour	**20.98**	26.31	28.48
PQCS	1 hour	15.98	**15.59**	16.66
	24-hour	**20.54**	27.41	26.05
	168-hour	**31.22**	36.92	39.13

The pinball loss and CRPS for KDEPF illustrate similar patterns as NDEPF. The scores rise sharply until the first 24-hour horizon, then they reach plateaus after about the 48-hour horizon. On the other hand, QCS and PQCS have an increasing trend until the 168-hour horizon for both electricity (Figure 4.15) and natural gas (Figure 4.16) datasets.

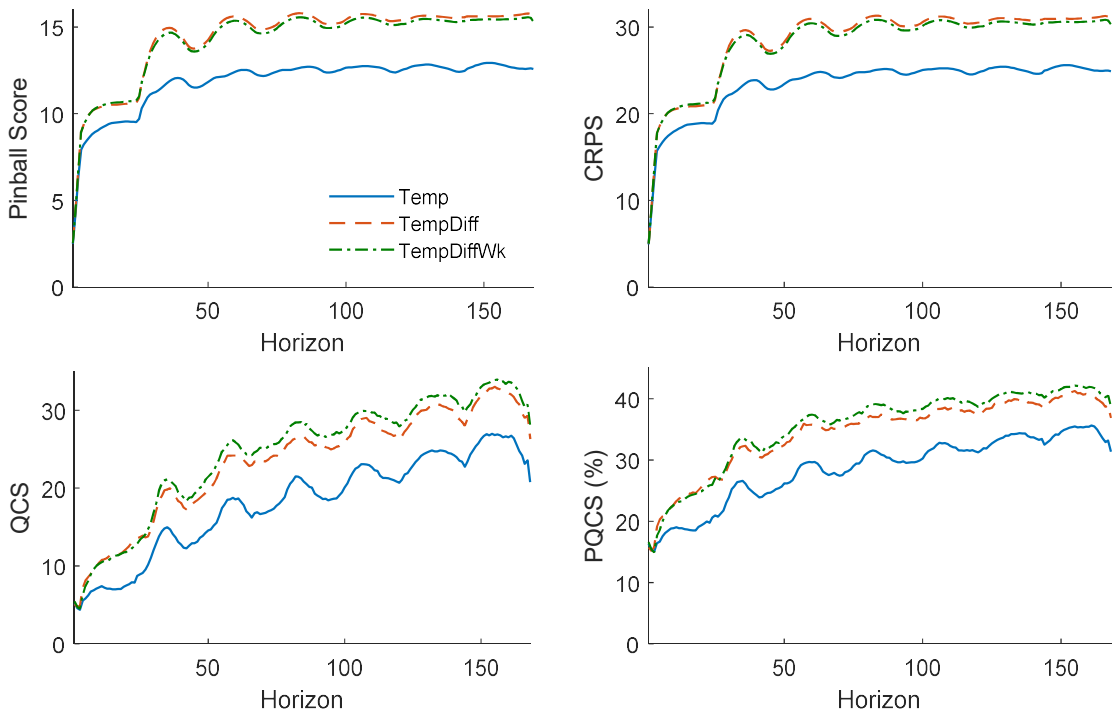

Figure 4.15: Two years average week-long hourly pinball scores, CRPS, QCS, and PQCS for three variants, KDEPF calculated from the electricity dataset.

The pinball score, CRPS, and QCS for the electricity dataset is better than the natural gas dataset. However, these three scores are dataset dependent. The PQCS demonstrates the opposite; natural gas probabilistic forecasts are better, although the natural gas point forecasts are poor compared to the electricity point forecasts. When point forecasts are relatively bad, then probabilistic forecasts show better performance. Hence, probabilistic forecasts compensate the shortcoming of poor point forecasts.

Table 4.8: Score comparison of three variants of the KDEPF using the natural gas dataset.

		KDEPF method (natural gas data)		
		Temperature	Temperature difference	Weekly temperature difference
Pinball score	1 hour	**136.28**	139.20	139.76
	24-hour	**606.40**	673.82	681.13
	168-hour	**752.03**	948.04	939.57
CRPS	1 hour	**269.97**	275.74	276.86
	24-hour	**1209.21**	1316.18	1351.03
	168-hour	**1499.88**	1940.35	1915.12
QCS	1 hour	**4.27**	8.95	9.56
	24-hour	10.19	**6.55**	6.68
	168-hour	20.41	**13.98**	14.44
PQCS	1 hour	**14.75**	20.96	21.56
	24-hour	23.21	**17.26**	18.05
	168-hour	31.76	26.16	**26.07**

In Figure 4.16, the temperature variant performed poorly compared to daily temperature difference and weekly average temperature difference using the QCS and PQCS. However, the outcome is opposite for the electricity dataset (Figure 4.15). This pattern matches with the result found for the NDEPF (Section 4.2.1). High variability of natural gas flow due to temperature swings might be the cause. Based on 20 years of hourly electricity and natural gas usage data, the highest electricity demand is approximately four times the lowest demand, whereas the highest natural gas flow is 49 times the lowest flow. Thus, the natural gas flow is more susceptible to the temperature changes than electricity demand.

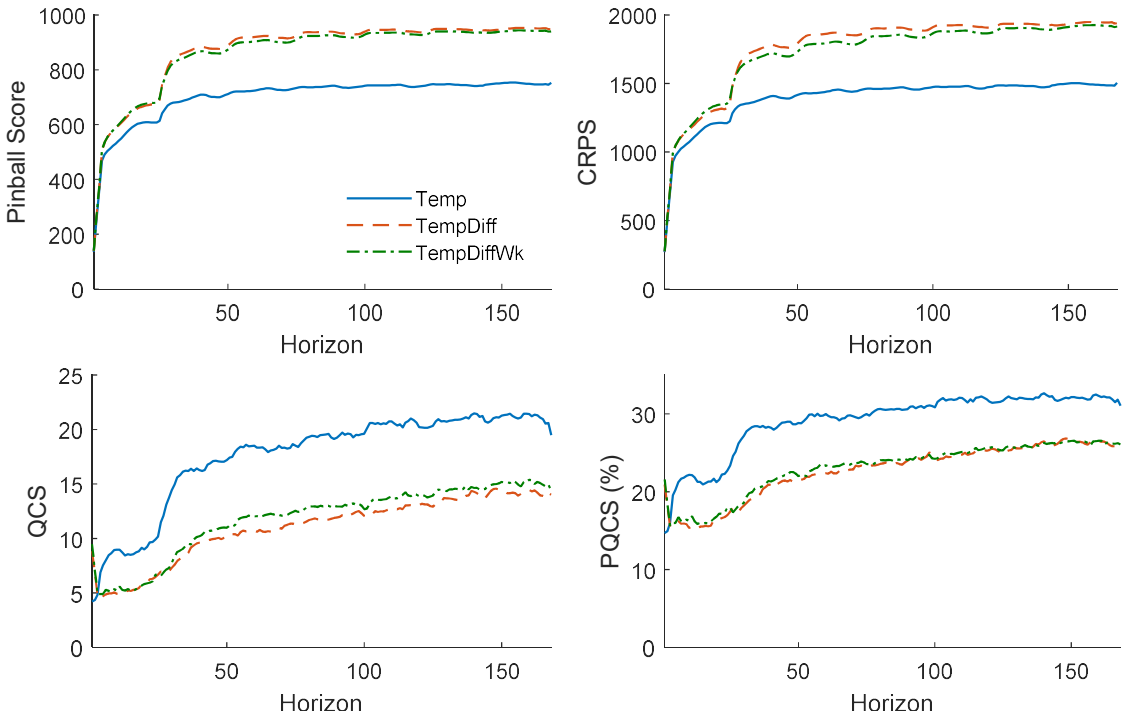

Figure 4.16: Two years average 168-hour horizon pinball scores, CRPS, QCS, and PQCS for three variants, KDEPF calculated from the natural gas dataset.

The next section presents more comparisons between the three different variants using the JDTPF method.

4.2.3 Performance Analysis of the JDTPF Method

This section presents week long hourly probabilistic forecasting scores using the JDTPF. Comparison among three different variants used in this work are presented in Table 4.9, Figure 4.17, Table 4.10, and Figure 4.18. Two different datasets are used: electricity and natural gas. Point forecasting errors generated from the MLR3 method are used to calculate JDTPF (see Section 4.1).

Table 4.9: Score comparison of three variants of the JDTPF using the electricity dataset.

		JDTPF method (electricity data)		
		Temperature	Temperature difference	Weekly temperature difference
Pinball score	1 hour	**2.49**	2.53	2.53
	24-hour	**9.54**	10.62	10.75
	168-hour	**12.62**	15.50	15.31
CRPS	1 hour	**4.93**	5.01	5.01
	24-hour	**18.90**	21.04	21.30
	168-hour	**25.01**	30.72	30.34
QCS	1 hour	**4.22**	4.34	4.60
	24-hour	**8.57**	13.21	12.49
	168-hour	**21.26**	26.37	28.23
PQCS	1 hour	14.33	**14.24**	14.98
	24-hour	**20.16**	27.01	25.58
	168-hour	**31.19**	36.69	38.78

Table 4.9 shows 1 hour, 1 day, and 1-week horizon probabilistic forecasting scores for the electricity dataset, where the temperature variant has done better than other two variants (similar to the NDEPF and the KDEPF). Figure 4.17 demonstrates 1 to 168-hour horizon probabilistic forecasting scores using the JDTPF. Scores found in this section are similar to the KDEPF method (Section 4.2.2).

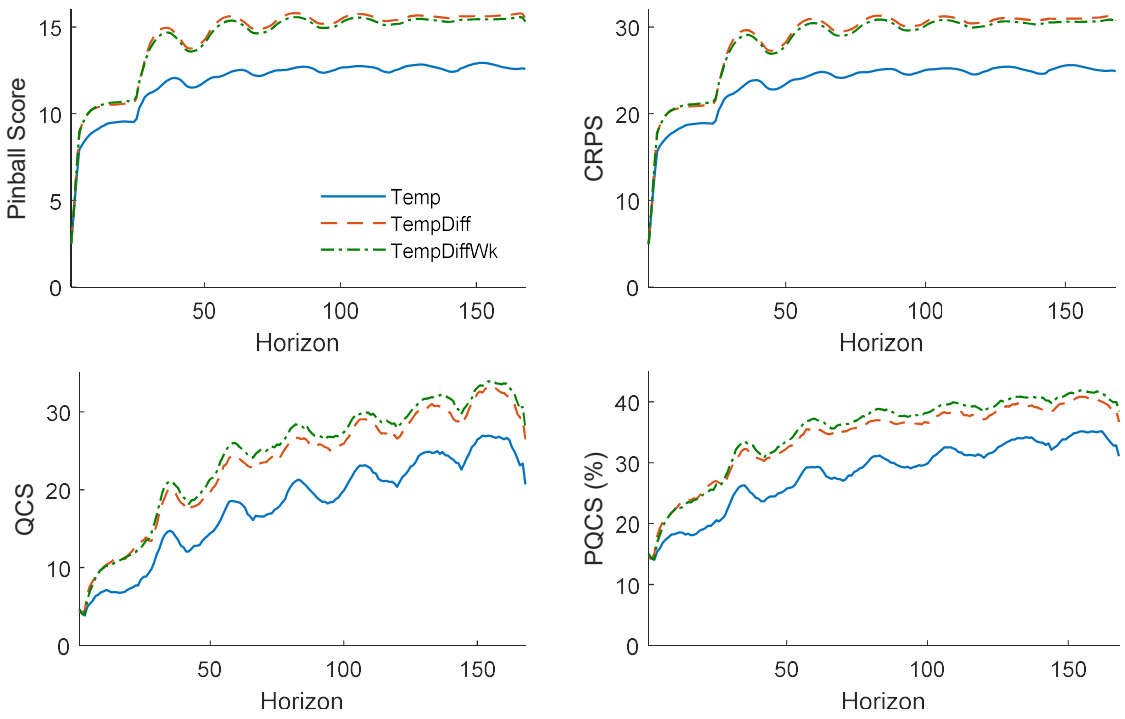

Figure 4.17: Two years average week-long hourly pinball scores, CRPS, QCS, and PQCS for three variants, JDTPF calculated from the electricity dataset.

Table 4.10 shows 1 hour, 1 day, and 1 week horizon natural gas probabilistic forecasting scores (pinball scores, CRPS, QCS, and PQCS). Three variants are used (similar to the Section 4.2.2) to generate probabilistic forecasts using the JDTPF. The temperature binning process is a better variant than the other two variants according to the pinball score and CRPS. However, daily temperature difference (Equation (3.3)) and weekly average temperature difference (Equation (3.4)) variants outperform the temperature variant according to the QCS and PQCS (Figure 4.18).

Table 4.10: Score comparison of three variants of the JDTPF using the natural dataset.

		JDTPF method (natural gas data)		
		Temperature	Temperature difference	Weekly temperature difference
Pinball score	1 hour	**136.25**	139.15	139.72
	24-hour	**606.49**	674.69	681.39
	168-hour	**752.29**	948.44	940.32
CRPS	1 hour	**269.96**	275.71	276.83
	24-hour	**1208.69**	1315.16	1348.48
	168-hour	**1496.56**	1934.49	1906.49
QCS	1 hour	**3.93**	8.59	9.18
	24-hour	10.02	6.43	**6.29**
	168-hour	20.34	**13.83**	14.20
PQCS	1 hour	**14.07**	20.76	21.23
	24-hour	23.03	**17.20**	17.41
	168-hour	31.38	26.02	**25.77**

The JDTPF method performed slightly better than the KDEPF for both electricity and natural gas datasets. However, the improvement is not statistically significant at the 5% confidence level (using the Kolmogorov-Smirnov test). More comparisons between three probabilistic forecasting methods are presented in the next section. Every probabilistic forecasting method has three variants. Hence, there are nine different ways to generate probabilistic forecasts. However, only the temperature variant is considered while comparing between different probabilistic forecasting methods in the next section for simplicity. Results obtained from the three variants are similar. In this section, the temperature variant outperformed the other two variants most of the time, which is the main reason to select the temperature variant for further analysis in the next section.

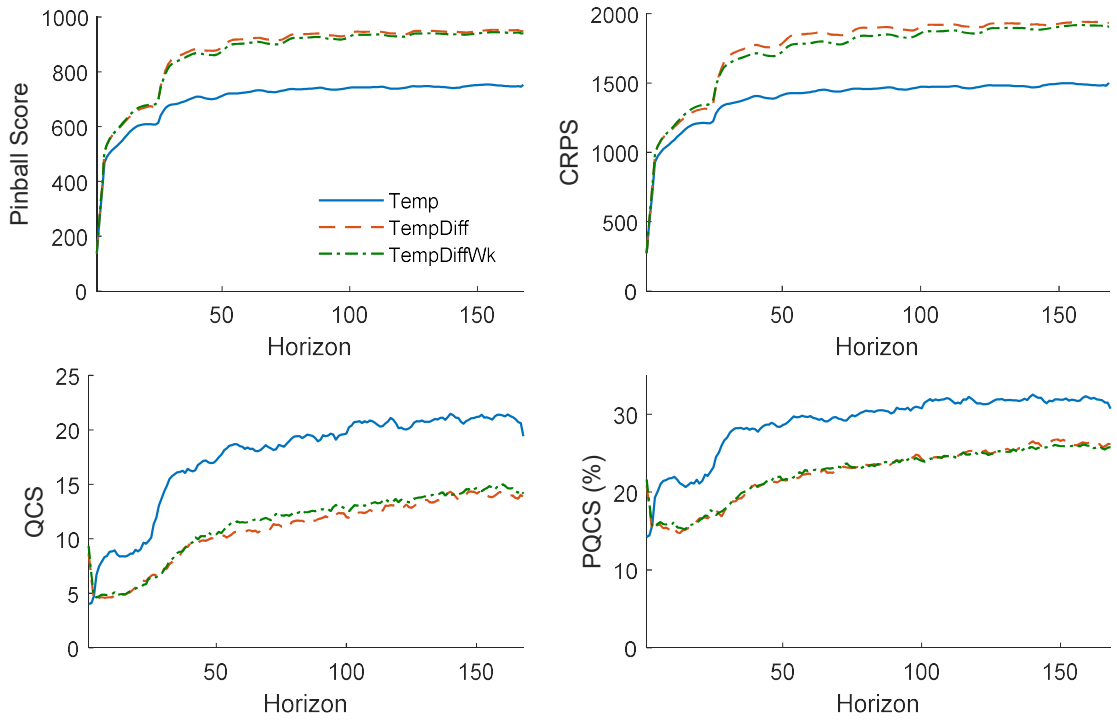

Figure 4.18: Two years average week-long hourly pinball scores, CRPS, QCS, and PQCS for three variants, JDTPF calculated from the natural gas dataset.

4.2.4 Comparisons Among NDEPF, KDEPF, and JDTPF Methods

This section compares three probabilistic forecasting methods explained in Sections 3.2, 3.3, and 3.4. Only the temperature variant (binning process) is used for simplicity. The other two variants produce similar results. Figure 4.19 illustrates the performance of NDEPF, KDEPF, and JDTPF (generated from the electricity dataset) based on the new evaluation technique, graphical calibration measure (GCM) (see Section 3.5).

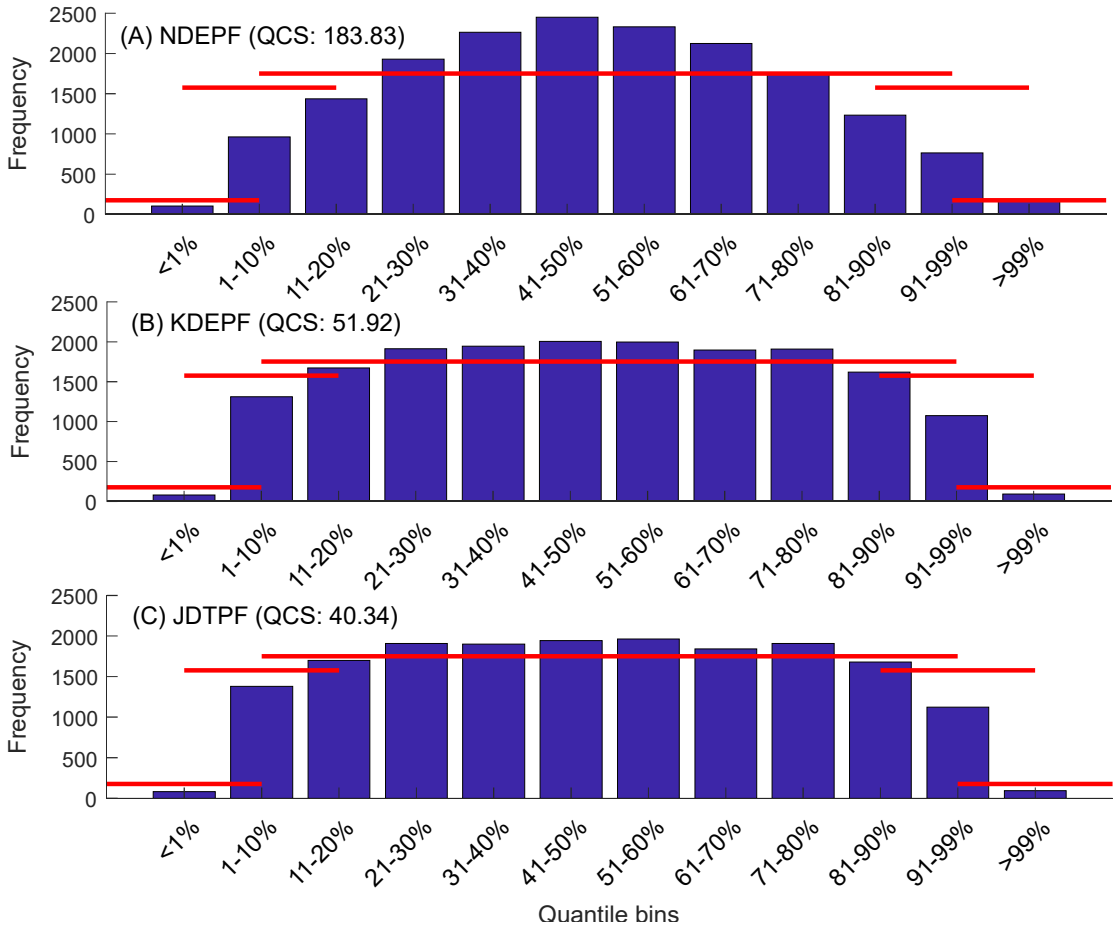

Figure 4.19: Performance analysis of (A) NDEPF, (B) KDEPF, and (C) JDTPF methods using GCM for one-hour horizon (electricity dataset).

The NDEPF method produces sharper CDFs than the KDEP and JDTPF methods, which is the reason for the big hump in Figure 4.19. The bar graphs show the frequency of observed demands within each quantile bin. The red line in the bar chart shows the expected height of each bar. Small difference between the frequencies of observed value and expected frequency (red line) indicates better probabilistic forecasts and vice versa. It is noticeable from Figure 4.19 that the KDEPF and JDTPF methods outperformed the NDEPF. The KDEPF and JDTPF look very competitive. QCS or PQCS can be used to break the tie. In this example, the JDTPF method (QCS: 31.90) performed better than the

KDEPF method (QCS: 43.92) based on the QCS. Table 4.11 shows 1 hour, 1 day, and 1 week horizon scores of the three probabilistic forecasting methods.

Table 4.11: Score comparison of three probabilistic forecasting methods (NDEPF, KDEPF, and JDTPF) using the electricity dataset.

		Probabilistic forecasting methods (electricity data)		
		NDEPF	KDEPF	JDTPF
Pinball score	1 hour	2.53	**2.49**	**2.49**
	24-hour	9.56	**9.54**	**9.54**
	168-hour	**12.62**	**12.62**	**12.62**
CRPS	1 hour	5.00	**4.93**	**4.93**
	24-hour	18.93	**18.90**	**18.90**
	168-hour	**25.00**	**25.00**	25.01
QCS	1 hour	13.81	4.91	**4.22**
	24-hour	11.06	8.72	**8.57**
	168-hour	21.60	**20.98**	21.26
PQCS	1 hour	29.16	15.98	**14.33**
	24-hour	23.49	20.54	**20.16**
	168-hour	32.45	31.22	**31.19**

Figure 4.20 compares scores for the three probabilistic forecasting methods (NDEPF, KDEPF, and JDTPF) for one to 168-hour horizons using the electricity dataset. The pinball score and the CRPS are the same for three probabilistic forecasting methods. The zoomed version of the Figure 4.20 (not included in this dissertation) shows tiny differences among different scores calculated from the three probabilistic forecasting methods, although the difference is not statistically significant at a 5% significance level (using the Kolmogorov-Smirnov test). However, KDEPF and JDTPF outperform NDEPF based on the QCS and the PQCS for each of 168 horizons.

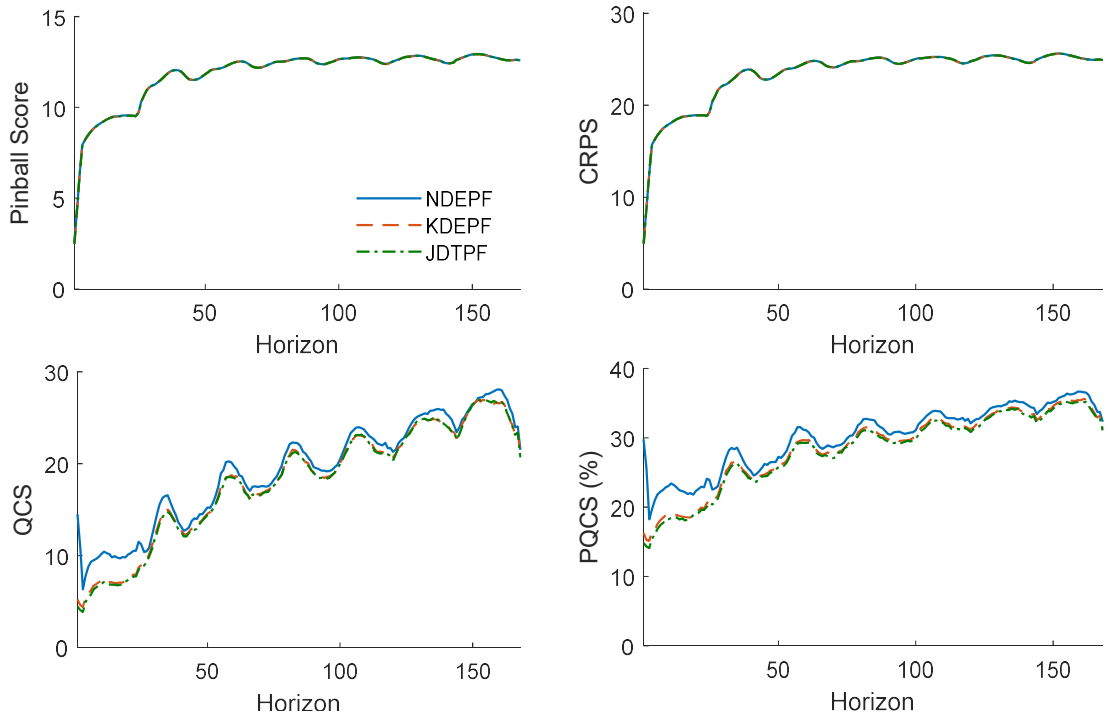

Figure 4.20: Week-long hourly score comparison between the three probabilistic forecasting methods (NDEPF, KDEPF, and JDTPF) using the electricity dataset.

Figure 4.21 compares the three probabilistic forecasting methods using the natural gas dataset. Like the electricity dataset, the KDEPF and JDTPF significantly outperform NDEPF. For the natural gas dataset, the KDEPF and JDTPF methods even performed better than the electricity dataset because there is more variability in the natural gas dataset. The NDEPF performed worse than the electricity dataset for the same reason.

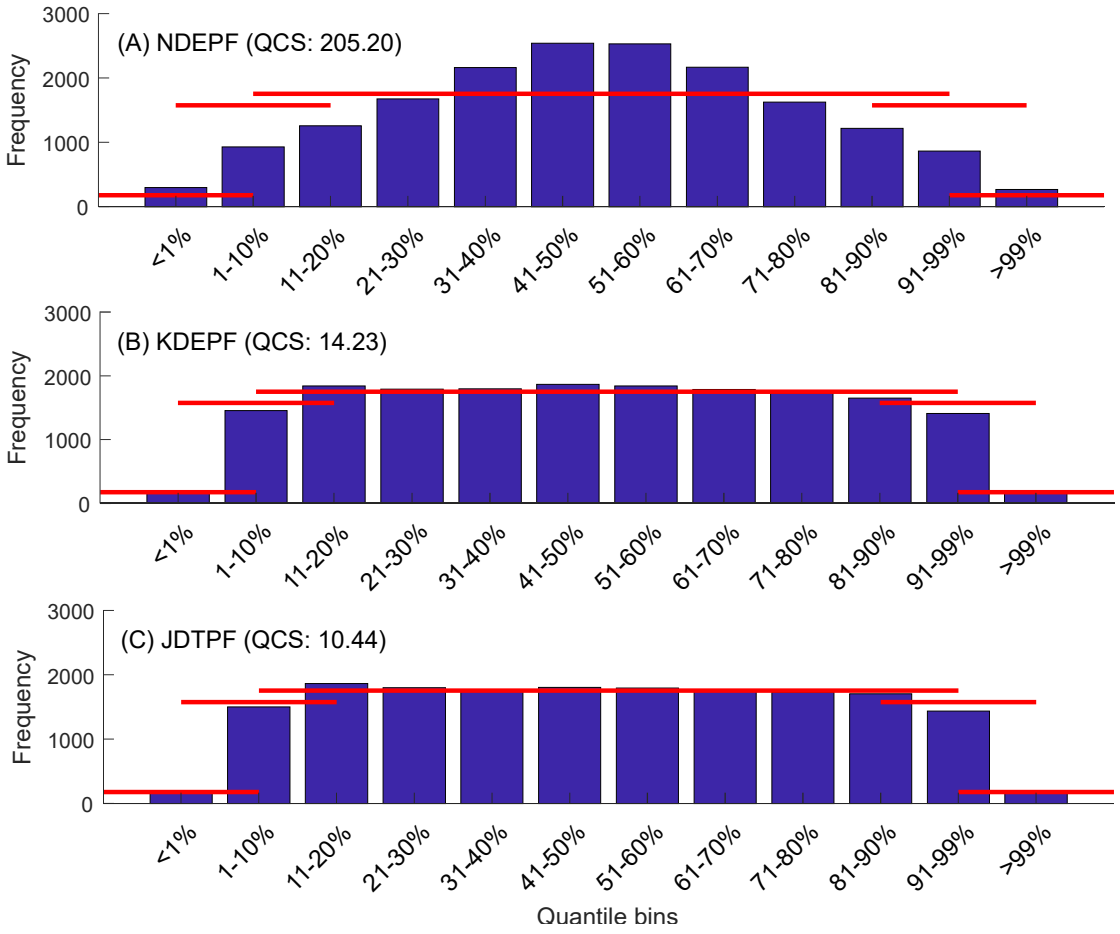

Figure 4.21: Performance analysis of (A) NDEPF (B) KDEPF, and (C) JDTPF methods using GCM for horizon one (natural gas dataset).

Table 4.12 displays 1 hour, 1 day, and 1 week horizon probabilistic forecasting scores for the natural gas dataset. Bold fonts indicate the best performance between three probabilistic forecasting methods. The JDEPF outperformed the other two probabilistic forecasting methods most of the time. However, the score difference between the KDEPF and JDTPF is not significant with a 5% significance level (using the Kolmogorov-Smirnov test). The running time of JDTPF is three times faster than the NDEPF, and around 1500 times faster than the KDEPF (Table 4.13) for both electricity and natural gas datasets.

Table 4.12: Score comparison of three probabilistic forecasting methods (NDEPF, KDEPF, and JDTPF) using the natural dataset.

		Probabilistic forecasting methods (natural gas data)		
		NDEPF	KDEPF	JDTPF
Pinball score	1 hour	138.50	136.28	**136.25**
	24-hour	606.47	**606.40**	606.49
	168-hour	**751.24**	752.03	752.29
CRPS	1 hour	274.06	269.97	**269.96**
	24-hour	**1193.82**	1209.21	1208.69
	168-hour	**1491.22**	1499.88	1496.56
QCS	1 hour	17.21	4.27	**3.93**
	24-hour	11.83	10.19	**10.02**
	168-hour	20.75	20.41	**20.34**
PQCS	1 hour	32.24	14.75	**14.07**
	24-hour	25.23	23.21	**23.03**
	168-hour	32.00	31.76	**31.38**

Table 4.13 compares the average running time of the three probabilistic forecasting methods. MATLAB 2017a software running on a Windows 7 64-bit machine with Intel Core i5 processor Dual Core 3.40 GHz and 8 GB RAM is used in this experiment. The JDTPF method takes less than a second to generate 17,520 probabilistic forecasts (2 years), whereas the NDEPF method takes less than three minutes, and the KDEPF method takes around half an hour (excluding training time).

Table 4.13: Running time of one horizon probabilistic forecsts using different methods.

		Probabilistic forecasting methods		
		NDEPF	KDEPF	JDTPF
Running time	Electricity dataset	~ 1.37 sec	~ 782.31 sec	~ 0.46 sec
	Natural gas dataset	~ 1.51 sec	~ 710.06 sec	~ 0.50 sec

Figure 4.22 shows the score comparison between the three probabilistic forecasting methods for the 168-hour horizon calculated from the natural gas dataset. The pinball score and CRPS are poor compared to the electricity dataset because the natural gas dataset contains more variations. However, the QCS and PQCS are better compared to the electricity dataset. The NDEPF performed worst among the three probabilistic forecasting methods.

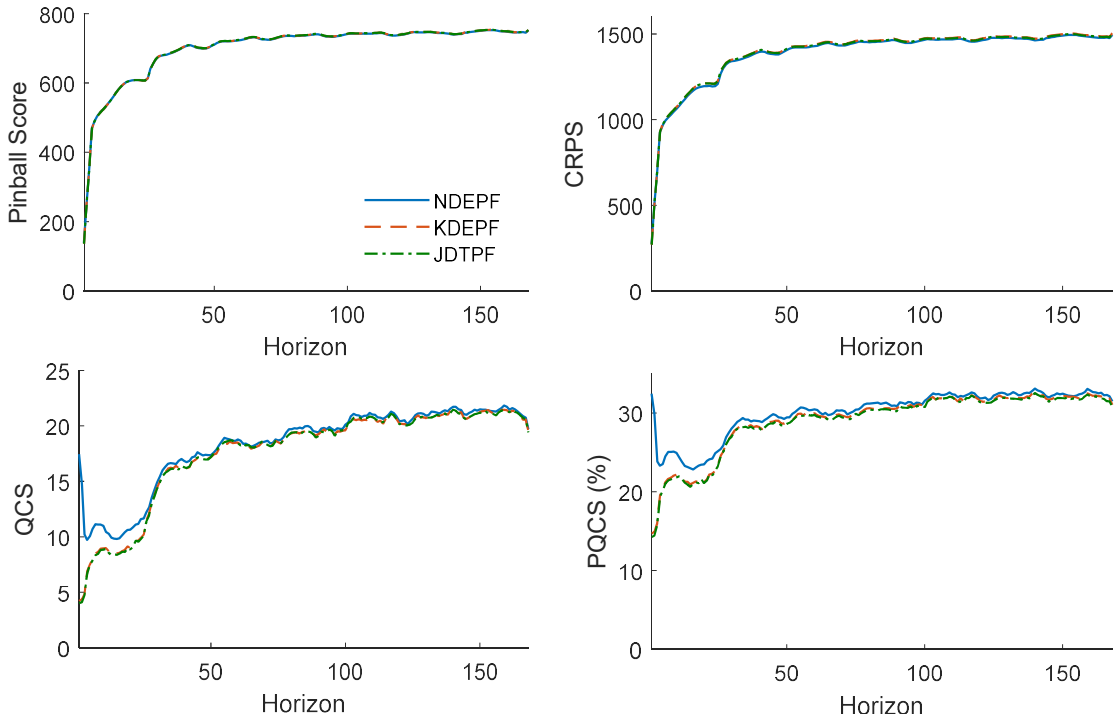

Figure 4.22: Week-long hourly score comparison between three probabilistic forecasting methods (NDEPF, KDEPF, and JDTPF) using the natural gas dataset.

Perfect weather data is used to generate point forecasts as well as probabilistic forecasts for this section. However, forecasted weather will be used in practice to generate probabilistic forecasts. The next section presents sample probabilistic forecasts and their scores generated from historical forecasted weather.

4.3 Probabilistic Forecast Using Forecasted Weather

This section shows experimental results of probabilistic forecasts using forecasted weather. Only the natural gas dataset is used in this experiment. The electricity dataset demonstrates similar outcomes. Twelve years of archived weather forecasts are collected from the GasDay^{TM} repository. Figure 4.23 demonstrates hours of the day (0-23), when forecasts are made. In the natural gas industry, forecasts are generally made before 10 A.M. (between 5 and 9 A.M.) most of the time to set hourly maximum flow limits (Figure 1.1) and natural gas purchase decisions for the coming 24-hour horizon. After 10 A.M., forecasts are made for monitoring the situation when the natural gas flow is very close to the maximum flow limit. For example, see the large number of forecasts made between 9 and 10 P.M. (Figure 4.23) over the long period of time. The forecasted weather dataset contains 857 days of multiple forecasts, and 467 days have no forecast (missing data). Hence, the quality of the dataset is poor compared to the dataset used in the earlier section. However, all three probabilistic forecasting methods (NDEPF, KDEPF, and JDTPF) are able to generate reasonable probabilistic forecasts.

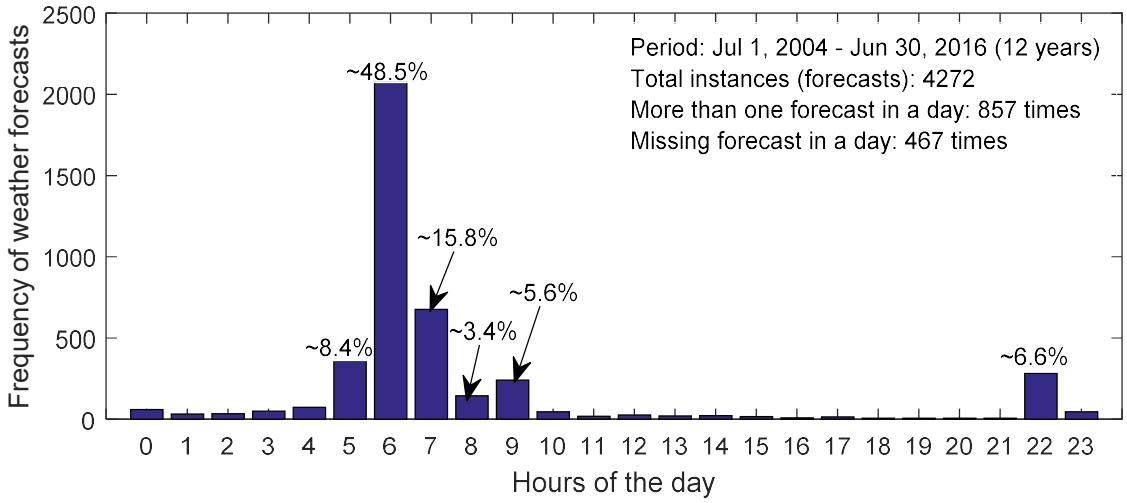

Figure 4.23: Hours of the day, when point forecasts are made.

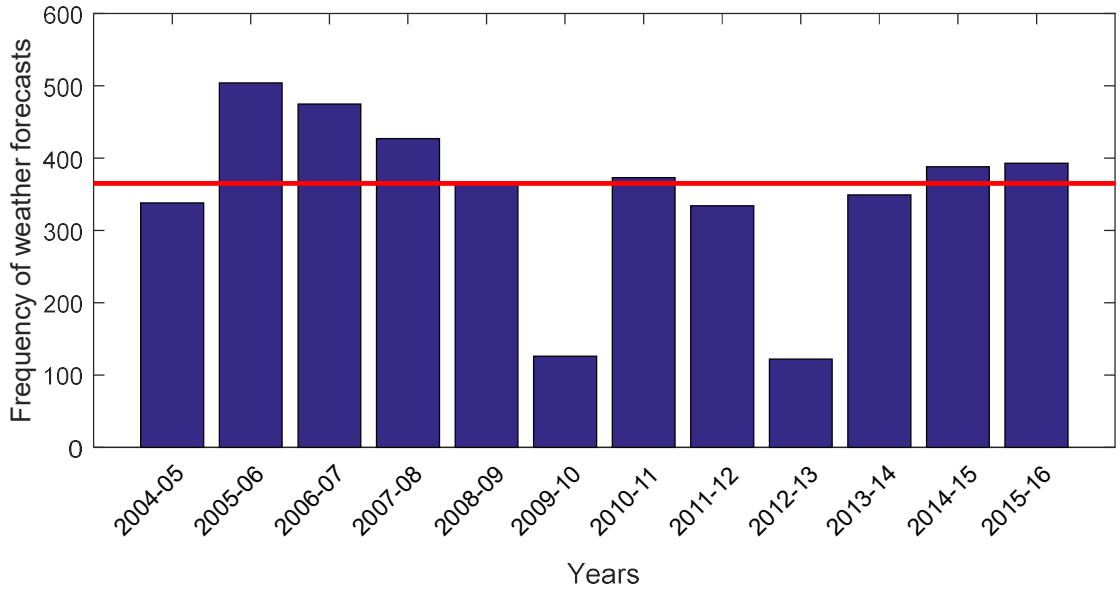

Figure 4.24: Yearly frequency of point forecasts (Jul 1, 2004- Jun 30, 2016)

Figure 4.24 shows the frequency of yearly forecast data available in the dataset. The red line indicates on average at least one forecast is available per day for a year. This experiment assumes that forecasts made at various times have the same impact on the

natural gas flow, which is not true in practice. Missing values are removed from the dataset before training forecasting models. No weather forecast was available in the GasDay^{TM} repository from Oct 9, 2009 to Jun 11, 2010 and Oct 30, 2012 to Sep 30, 2013 (see two shortest bars in Figure 4.24). A similar data processing technique explained in Sections 4.1 and 4.2 are used to create training and testing subsets for point (Table 4.1) and probabilistic forecasts (Table 4.4), respectively.

Figure 4.25 shows an example day-long hourly probabilistic forecasts generated from forecasted weather. In this figure, the forecasted CDFs are sharper than actual weather generated forecasts (Section 4.2). Fewer data points are available in the forecasted weather dataset than in the previous section, which might be the cause of the extra sharpness. It is possible to capture more variability when more data points are available in the dataset. Figure 4.26 demonstrates another forecasted weather day-long hourly probabilistic forecasts, which has less sharp forecasted CDFs than the one in Figure 4.25. The sharpness of forecasted CDFs is different for different seasons, which is noticeable in the forecasted weather generated probabilistic forecasts. However, the time horizon variability is less pronounced in the forecasted weather generated probabilistic forecasts than actual weather generated probabilistic forecasts because very few forecasts are available between 10 A.M. and 9 P.M. (Figure 4.23). It is still possible to generate probabilistic forecasts from the bad forecasted weather dataset, which validates the credibility of the three probabilistic forecasting methods (NDEPF, KDEPF, and JDTPF). The JDTPF is used to demonstrate example probabilistic forecasts for this section; other methods provide similar results.

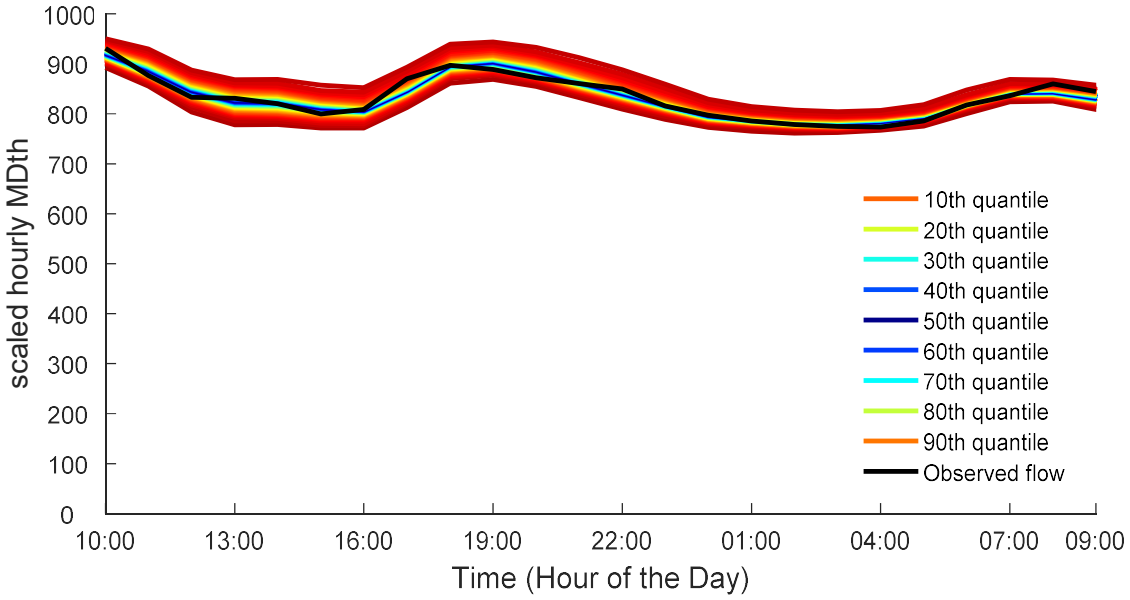

Figure 4.25: An example day-long hourly probabilistic forecasts generated form forecasted weather (Date: Jan 16, 2015)

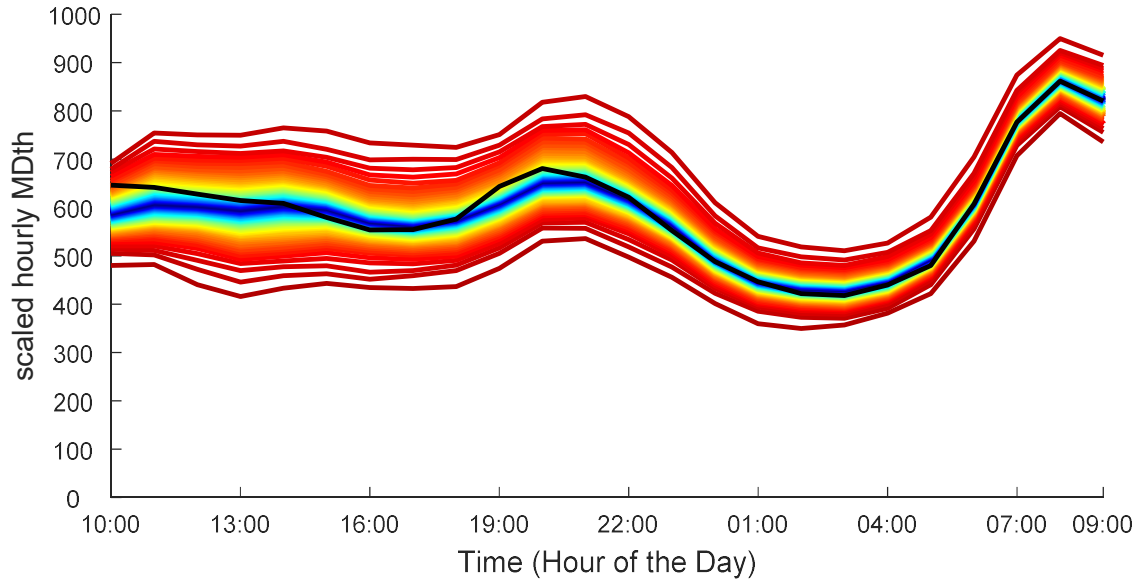

Figure 4.26: An example day-long hourly probabilistic forecasts generated from forecasted weather data (Date: Sep 29, 2014)

Figure 4.27 shows the performance of forecasted weather generated probabilistic forecast using the graphical calibration measure (GCM) for horizon one. Probabilistic forecasts are below the 1% quantile about 5% of the time (expectation was 1%), which is not good. Overall, forecasts are biased towards the left. However, the three probabilistic forecasting methods explained Sections 3.2, 3.3, and 3.4 are good enough to generate probabilistic forecasts from even a bad dataset.

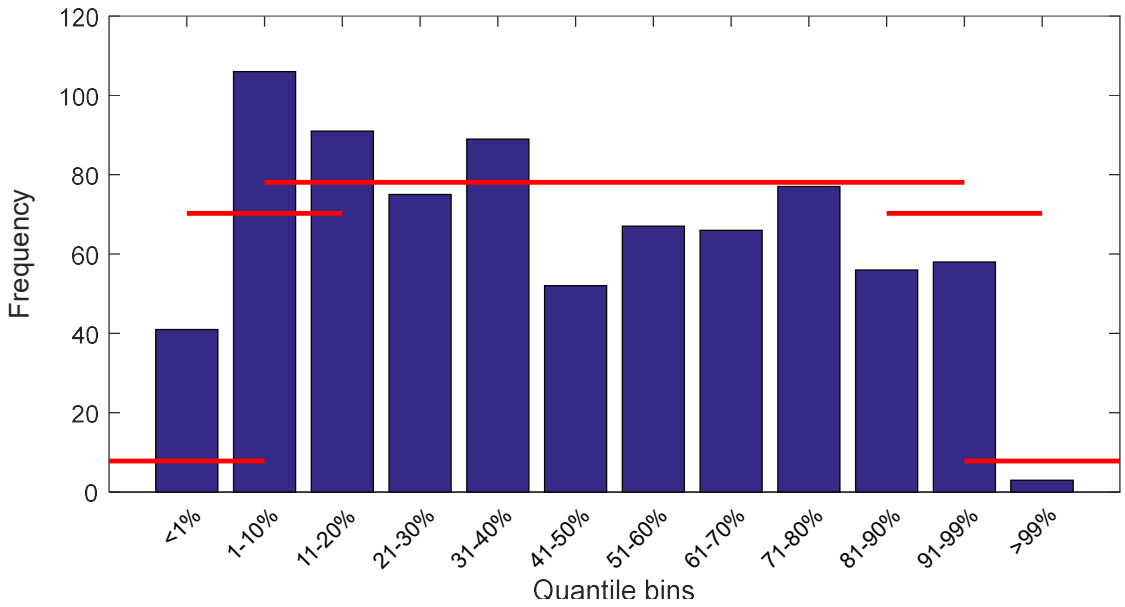

Figure 4.27: Assessment of forecasted weather generated probabilistic forecasts using the graphical calibration measure (PQCS: 18.87)

Probabilistic forecasts are less difficult to make from forecasted weather compared to point forecasts. It is shown in the previous section that the probabilistic forecast performs even better in the sense of capturing uncertainty, when forecasting elements (natural gas or electricity) are more variable.

Forecasting unusual days (such as large temperature swings, colder than normal days, warmer than normal days, etc.) is comparatively more challenging and more important to forecast well than normal days in the natural gas industry. Probabilistic forecasts are found useful and add value with point forecasts during those difficult forecasting days. The next section presents the performance of probabilistic forecasts during unusual days.

4.4 Unusual Days Analysis for Probabilistic Forecasts

In the natural gas industry, it is important to forecast well during unusual days because unusual days are difficult to predict. Thus, there are more chances to pay larger penalties for unusual days if special attention is not given. Vitullo (2011) identified 10 classes of unusual days [135] for the natural gas industry (see Table 4.14). This section compares the performance of three probabilistic forecasting methods (NDEPF, KDEPF, and JDTPF) during the top 5% of the unusual days with forecasts for all days.

Table 4.14: Unusual day types for natural gas forecasts [135]

Unusual days			
1	Coldest days	2	Colder than normal days
3	Warmer than normal days	4	Windiest heating days
5	Colder than yesterday	6	Warmer than yesterday
7	First cold days	8	First warm days
9	High humidity heating days	10	Low humidity heating days

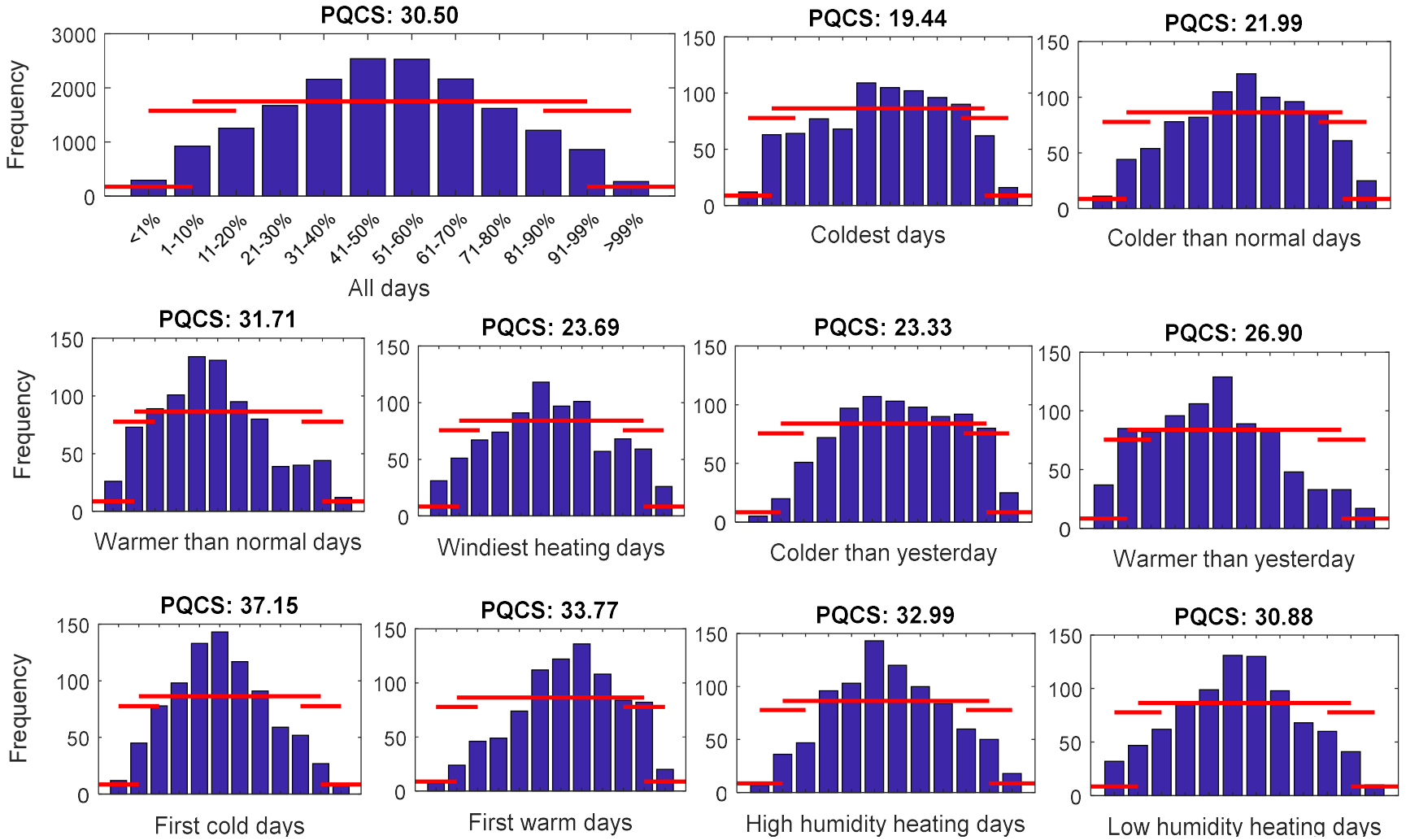

Figure 4.28: Performance of the NDEPF during unusual days

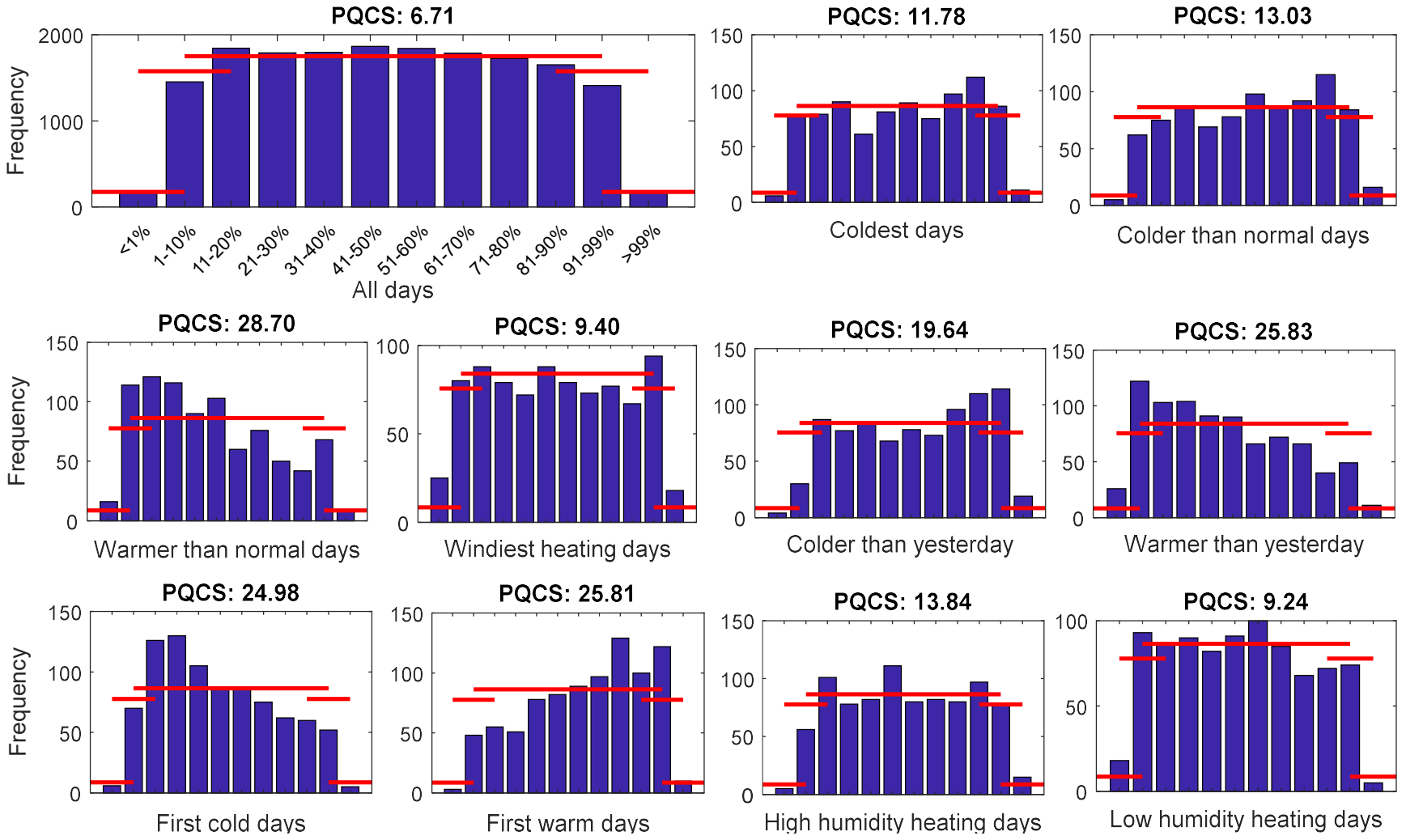

Figure 4.29: Performance of the KDEPF during unusual days

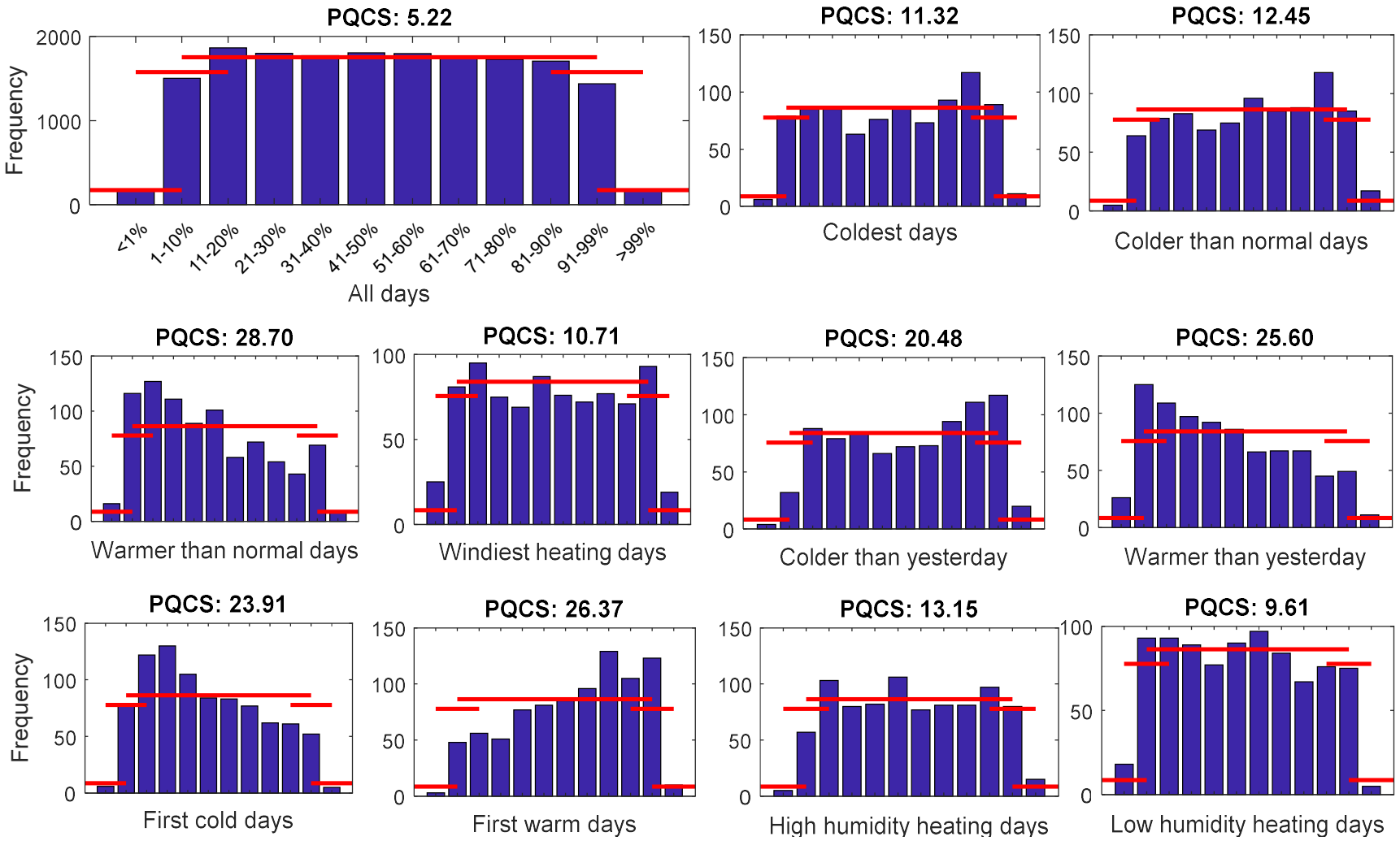

Figure 4.30: Performance analysis of the JDTPF during unusual days

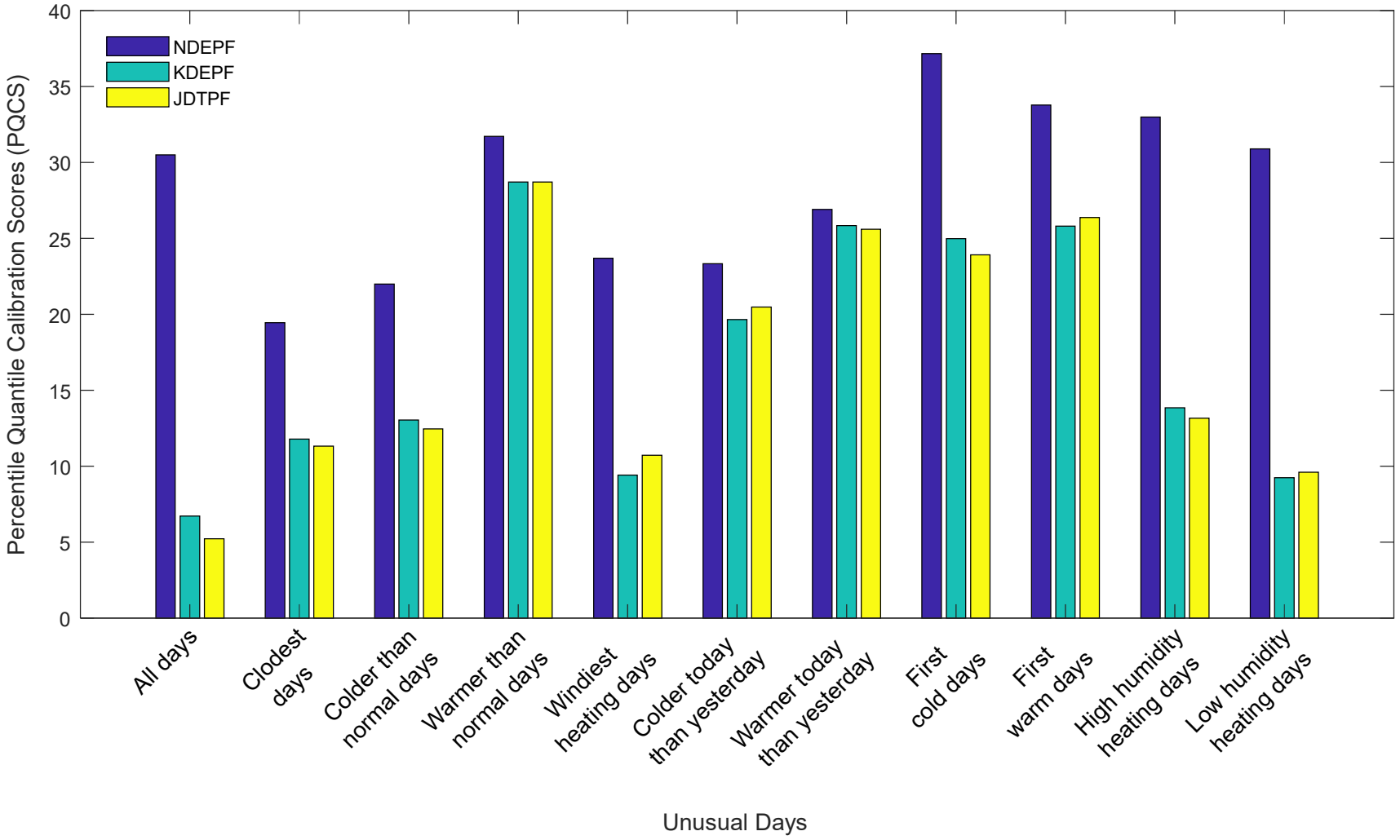

Figure 4.31: Unusual days PQCS comparison among the three probabilistic forecasting methods

Unusual days are usually the top 5% most difficult days to forecast. Hence, each unusual day dataset is 20 times smaller (~ 864 probabilistic forecasts) compared to all days (17,520 probabilistic forecasts). Thus, only PQCS is applicable to compare between all days and unusual days, because the PQCS is data independent. The existing unusual day algorithm [135] is designed based on daily data, so it may not work well for hourly data. However, it is the only currently available algorithm to find unusual events from weather data, and finding unusual hours is out of scope for this dissertation. Thus, unusual days are converted to unusual hours by picking all 24 hours of an unusual day.

Figure 4.28 compares the performance of NDEPF method during the 10 kinds of unusual days with its performance on all days. Unusual day forecasts also are concentrated in the middle (near to the 50^{th} quantile) like all days. Coldest days performed reasonable compared to all days, while first cold days produce worse PQCS among all unusual day types. Five out of 10 unusual days performed better than all days for the NDEPF.

Figure 4.29 shows a comparison between performance on unusual days and performance on all days based on the PQCS, where performance on all days outperformed performance on all 10 unusual days. Performance of probabilistic forecasts on warmer than normal days, warmer than yesterday and first cold days are skewed on the left, which indicates under-forecast for those unusual days. On the other hand, evidence of over-forecasting is found during first warm days. The JDTPF method shows similar performance for unusual days on this same dataset.

Figure 4.30 shows the graphical calibration measure (GCM) of unusual days using the JDTPF method. The performance of probabilistic forecasts during unusual days are poor compared to performance on all days. Figure 4.31 compares the performance of the three probabilistic forecasting methods (NDEPF, KDEPF, and JDTPF) during unusual days. The KDEPF and the JDTPF outperformed NDEPF during unusual days as well as all days. The JDTPF performed better than the KDEPF during coldest days, colder than normal days, and first cold days. On the other hand, the KDEPF outperformed the JDTPF during windiest heating days, and colder than yesterday.

This chapter showed the results from the MLR3 point forecasting method described in Section 4.1. The MLR3 point forecasting method is used to generate probabilistic forecasts in this work. The performance analysis of three probabilistic forecasting methods, NDEPF, KDEPF, and JDTPF is presented in subsections 4.2.1, 4.2.2, and 4.2.3, respectively. The performance of three probabilistic forecasting methods is compared using four scoring rules (pinball score, CRPS, QCS, and PQCS) in Subsection 4.2.4. In Section 4.3, forecasted weather is used to repeat the same experiments done in Section 4.2. Finally, in Section 4.4 unusual day (5% most difficult days to forecast) analysis is offered for three probabilistic forecasting methods. The next chapter summarizes the findings of this dissertation with some proposed future work.

CHAPTER 5

CONCLUSIONS AND RECOMMENDATION FOR FUTURE WORK

This chapter presents a summary of the contributions made in this dissertation and proposed methods to quantify forecast uncertainties in Section 5.1. Important research findings and observations are summarized in Section 5.2. Some ideas to improve the presented methods and further research are proposed in Section 5.3. This dissertation focuses on quantifying forecast uncertainties with a goal to provide a useful tool for natural gas controllers. From an extensive literature review in Chapter 2, no literary evidence of probabilistic forecast usage is found in the natural gas industry, although informal anecdotal evidence is known. The lack of useful probabilistic forecast measuring tools is one of the main reasons why probabilistic forecasts are not frequently used in the energy sector [37].

In this dissertation, several methods are implemented to generate probabilistic forecasts through historical point forecast error analysis. A new evaluation technique is used to assess probabilistic forecasts. The data flow diagram of generating probabilistic forecasts is shown in Figure 5.1. At first, historical weather and flow data are cleaned and detrended. Several training and testing datasets are created (see Table 4.1) to generate point forecasts. The point forecast errors are analyzed to create a probabilistic forecasting engine, which can generate week-long hourly probabilistic forecasts. Finally, probabilistic forecasts are evaluated using a newly developed graphical score, GCM. Two energy demand datasets (natural gas and electricity) are use in this work as case studies.

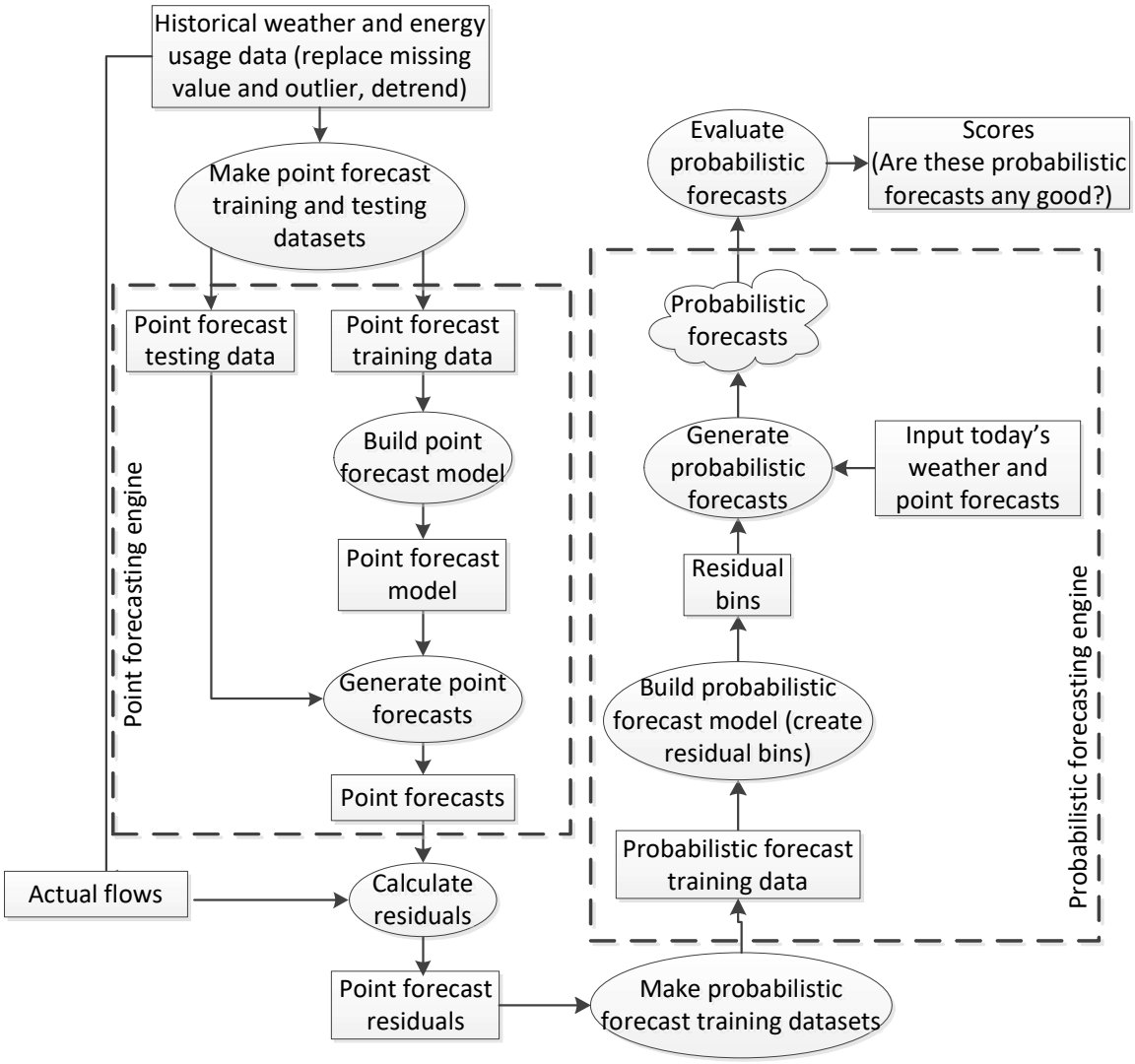

Figure 5.1: Data flow diagram of generating probabilistic forecast from a raw dataset

5.1 Contributions

This dissertation has three main contributions. First, three new methods: normal distribution estimator probabilistic forecast (NDEPF), kernel density estimator probabilistic forecast (KDEPF), and Johnson data transformation probabilistic forecast (JDTPF) are provided in Chapter 3 to produce probabilistic forecasts. Second, a graphical

probabilistic forecast measure technique, graphical calibration measure (GCM) is presented along with two new metrics, quantile calibration score (QCS) and percentile quantile calibration score (PQCS). Finally, probabilistic forecasts are applied in the natural gas industry to solve a real-life problem.

The proposed probabilistic forecasting methods (in Chapter 3) are applied on electricity and natural gas datasets to produce probabilistic forecasts. At first, observed weather is used to generate probabilistic forecasts, then forecasted weather is used to repeat the experiment. A benchmark method, NDEPF, is compared with other probabilistic forecasting methods, KDEPF and JDTPF. Currently, the natural gas industry is using the standard deviation of historical forecast errors to find the maximum and the minimum bound of a given point forecast (interval forecast). The benchmark method is an improved version of the existing method. Four scoring rules, pinball loss, CRPS, QCS, and PQCS are used to compare three probabilistic forecasting methods. Based on PQCS, the KDEPF and JDTPF performed around 13% - 17% better than the benchmark, NDEPF for horizon one. Experimental results show that the KDEPF and JDTPF outperform the benchmark method, NDEPF for one to 168 hour (1 week) horizons. The three probabilistic forecasting methods can generate credible probabilistic forecasts even from bad datasets (such as forecasted weather). The unusual day (top 5% difficult days to forecast) study demonstrates that, the probabilistic forecasts are more useful during bitter cold days, big temperature swings, and shoulder months (November, March). The KDEPF and JDTPF method outperform the benchmark, NDEPF during all unusual days as well as normal days. Overall, the JDTPF performs slightly better than the KDEPF.

However, the improvement is not statistically significant at the 5% level of significance. The running time of the JDTPF is around 1500 times faster than the KDEPF, and three times faster than the NDEPF. The next section summarizes important research findings and observations from this work.

5.2 Important Research Findings and Observations

Experimental results show that all three probabilistic forecasting methods (NDEPF, KDEPF, and JDTPF) generate credible probabilistic forecasts, even from a bad dataset (see Section 4.3). Probabilistic forecast evaluation technique, GCM is helpful to diagnose point forecasts during unusual days (see Section 4.4). For example, the MLR3 method is under-forecasting during cold days and over-forecasting during warm days (see Figures 4.29 and 4.30).

The KDEPF and JDTPF outperformed NDEPF in all situations including all unusual days based on four scoring rules (see Section 4.2.4), which indicates that the normality assumption of the residual distribution degrades the performance of uncertainty quantification. Thus, it is recommended to use the KDEPF or the JDTPF method instead of NDEPF method. However, the JDTPF is about 100 times faster than the KDEPF (see Table 4.13). So, the JDTPF is recommended.

When point forecasts are poor (Figures 4.2 and 4.4), the performance of probabilistic forecasts is comparatively better (Figures 4.20 and 4.22). Thus, probabilistic forecasts can compensate the shortcomings of point forecasts. Probabilistic forecasts

provide more information (entire CDF) compared to a point forecast (50^{th} quantile of the CDF). Thus, it is a useful tool to the gas controller to make better decisions in crisis situations (see Figure 1.1). However, one should keep in mind while using probabilistic forecasts that 1% of the time the actual flow is expected to be more than the 99^{th} quantile, and 1% of the time the actual flow is anticipated to be lower than the 1^{st} quantile.

This dissertation offered textual, numeric, and graphical presentation of probabilistic forecasts. The graphs are found most useful to communicate probabilistic forecasts and its scoring rules with practitioners. Thus, an assortment of colors is used to demonstrate probabilistic forecasts. The graphical evaluation technique, GCM, is presented along with two numerical scores to assess the performance of probabilistic forecasts. It is important to present complex calculations in an uncomplicated way to facilitate communication. One of the main goals of this work to find out better ways to communicate probabilistic forecasts. Then next section provides some ideas to extend this work in future.

5.3 Recommendation for Future Work

This research can be improved upon by exploring multi-variate binning techniques. In this work, only one variant (temperature, daily temperature difference, or weekly temperature difference) is used to create residual bins. Two or three variants can be considered at a time to create residual bins. Also, new variants such as heating degree days (HDD) or cooling degree days (CDD) can be introduced to create residual bins in

the process of generating probabilistic forecasts. A multi-variate binning process adds extra complexity to the existing problem. Thus, the main challenge of this method is to find an efficient algorithm to train the probabilistic forecast engine.

In this work, residual bins are created and stored in memory for generating probabilistic forecasts. Creating residual bins on the fly (when it is required) can be an alternative approach to the existing residual binning process (Figure 3.5). For a given temperature (say 65^0F), an on-the-fly binning process would collect historical residuals of a certain temperature range (say 65 ± 2^0F, i.e., between 63^0F and 67^0F) and create a one-time-use residual CDF. In this approach, historical residuals corresponding to the given temperature will be in the middle, which may improve the quality of residual CDFs and possibly probabilistic forecasts. In addition, this approach may reduce training time of the probabilistic forecasting methods (especially KDEPF, see Table 4.13) significantly. On the other hand, testing time will increase a little bit. The main challenge of this method is to keep the testing time within a reasonable level.

Other point forecasting methods such artificial neural networks (ANN), GasDay's dynamic post processor [208] (DPP, ensemble of linear regression and ANN), or deep neural networks (DNN) can be used as point forecasts to generate probabilistic forecasts. ANN, DPP, and DNN are usually considered better alternatives to linear regression for generating point forecasts (Figure 5.2). However, the (positive or negative) effect of using improved point forecasting methods on probabilistic forecasts is unknown. Further research required to answer this question.

Brown et al.'s detrending algorithm [207] is used in this work to improve point forecasts. Smoothing [209] further improves point forecasts (Figure 5.2). However, the effect of smoothing techniques on probabilistic forecasts is unknown. During large temperature swings, forecasted CDFs for two consecutive horizons may be significantly different. In this work, two consecutive bins are overlapped to reduce discontinuity. In addition to the bin overlapping, a triangular or parabolic smoothing technique can be used on probabilistic forecasts to further reduce the effect of big temperature swing between two consecutive forecast horizons.

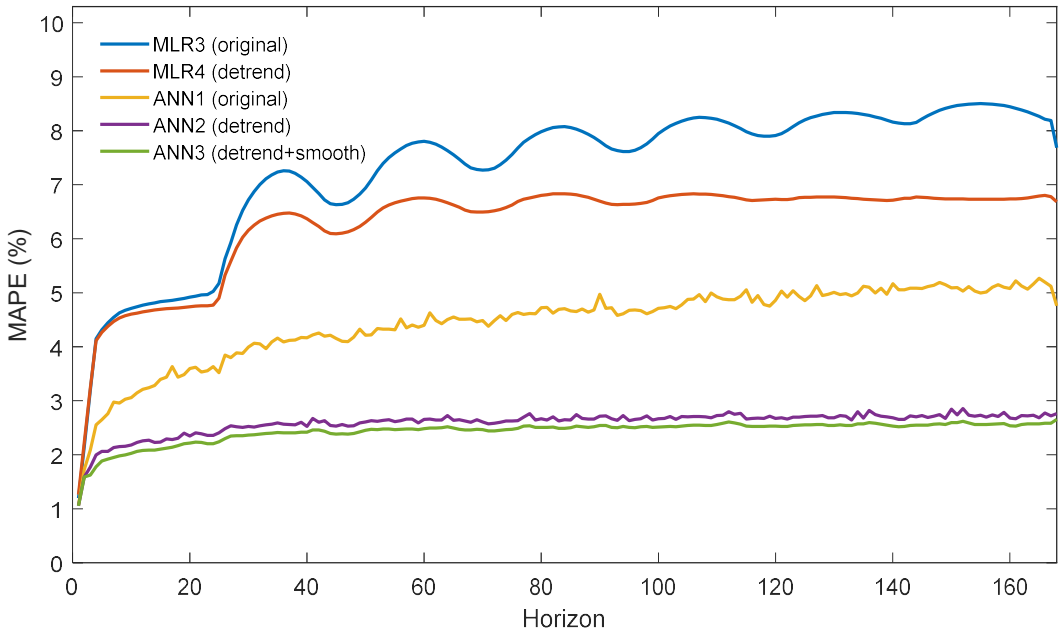

Figure 5.2: MAPE of different point forecasts (electricity dataset)

Instead of generating probabilistic forecasts from historical residuals, similar historical temperature scenarios (Figure 5.3) can be used to generate probabilistic forecasts. For instance, if the given temperature for generating probabilistic forecasts is

65^0F, then a search for historical point forecasts for similar temperature (say 65 ± 2^0F) will provide several point forecasts. In addition to the temperature, seasonality can be included to reduce the search space. Finally, a CDF (probabilistic forecast) can be generated from the historical point forecasts.

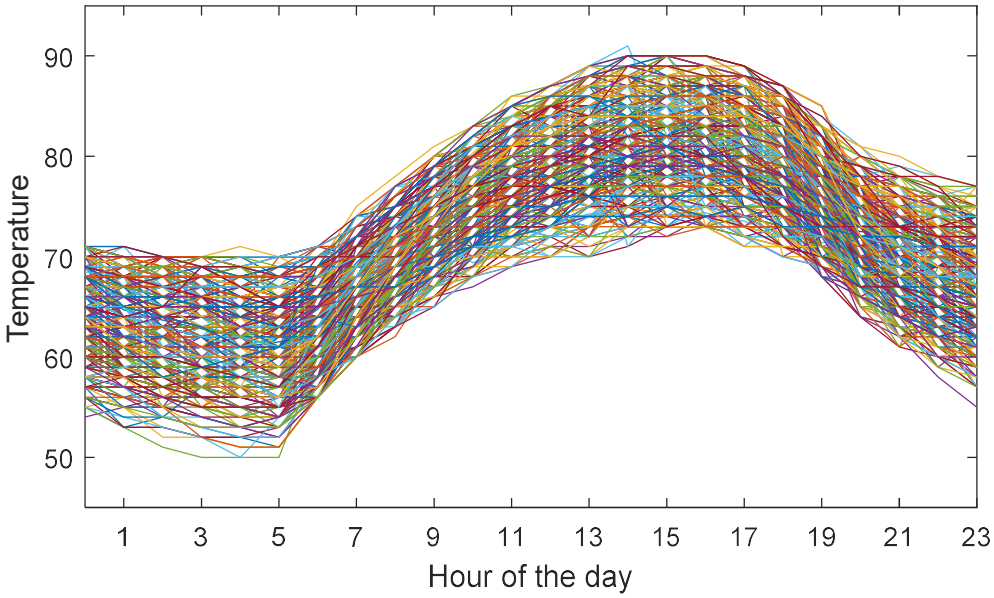

Figure 5.3: An example day long hourly temperature scenario generated from historical temperatures

In Section 4.4, unusual hour probabilistic forecasts are selected by the unusual days algorithm, because the unusual day for hourly data is not available. Unusual day analysis for hourly probabilistic forecasts might be more meaningful if an unusual hour algorithm were used. Research needs to be done to identify unusual hours and to explore their usefulness. The existing unusual day algorithm finds unusual days considering heating season, which is more appropriate for the natural gas industry, and less effective

for the electrical industry. A new algorithm is required to find unusual days/hours for the electricity industry.

It is well established that combining individual good point forecasts improves the accuracy of point forecast [140, 143, 208]. However, combining probabilistic forecast still a new area of research [37]. A naïve method (simple averaging) is applied to combine the three probabilistic forecasting methods presented in this dissertation. The combined probabilistic forecast could not outperform all individual probabilistic forecasts. More research is required to find a better way to combine probabilistic forecasts.

This dissertation focuses on quantifying forecast uncertainty in the energy industry. The same probabilistic forecasting methods (NDEPF, KDEPF, and JDTPF) and evaluation techniques (GCM, QCS, PQCS) presented in this dissertation can be used in other sectors such as health care, meteorology, and economics to quantify forecast uncertainties. For example, Amedu, in his master's essay [210] applied the NDEPF method to forecast mosquito age, which is helpful to identify deadly mosquitos carrying malaria. He also applied the newly introduced GCM technique to evaluate his probabilistic forecasts. A journal paper on Amedu's essay is in preparation [211].

BIBLIOGRAPHY

[1] T. Content, "Lab Learns to Predict Gas Demand: Marquette Training Venture Saves Utilities Millions," *Milwaukee Journal Sentinel,* 9 Jan 2010.

[2] T. Connery and T. Stephens, *PGL Industry Overview,* Marquette University, Wisconsin: Unpublished presentation slides, GasDay seminar, Apr, 2016.

[3] R. Weron, Modeling and Forecasting Electricity Loads and Prices: A Statistical Approach, West Sussex, England: John Wiley and Sons, 2006.

[4] G. Gross and F. D. Galiana, "Short-Term Load Forecasting," *Proceedings of the IEEE,* vol. 75, no. 12, pp. 1558-1573, 1987.

[5] M. Shahidehpour, H. Yamin and Z. Li, Market Operations in Electric Power Systems, New York: IEEE, 2002.

[6] S. Lakshminarayana and Anjul, "Smart Grid Technology and Applications," in *Power and Energy Systems: Towards Sustainable Energy (PESTSE)*, Bangalore, India, 2014.

[7] A. Y. Saber and G. K. Venayagamorthy, "Resource Scheduling Under Uncertainty in a Smart Grid With Renewable and Plug-in Vehicles," *IEEE Systems,* vol. 6, no. 1, pp. 103-109, 2012.

[8] E. Zio and T. Aven, "Uncertainties in Smart Grids Behavior and Modeling: What are the Risks and Vulnerabilities? How to Analyze Them?," *Energy Policy,* vol. 39, no. 10, pp. 6308-6320, 2011.

[9] "DSIRE (Database of State Incentives for Renewables and Efficiency)," NC Clean Energy Technology Center, 20 June 2016. [Online]. Available: http://www.dsireusa.org/. [Accessed 20 June 2016].

[10] The White House, Office of the Press Secretary, "FACT SHEET: U.S. Reports its 2025 Emissions Target to the UNFCCC," The White House, 31 March 2015. [Online]. Available: https://www.whitehouse.gov/the-press-office/2015/03/31/fact-sheet-us-reports-its-2025-emissions-target-unfccc. [Accessed 20 June 2016].

[11] P. Pinson, "Renewable Energy Forecasts Ought to be Probablistic," Center for Electric Power and Energy, 11 Aug 2015. [Online]. Available: https://www.youtube.com/watch?v=W65Hm2PbV5E&index=14&list=PLqNlQW--9bUQwJc0HvgoyFAW5rR9jC3lr. [Accessed 24 Jan 2017].

[12] U.S. Energy Information Administration, "Energy in Brief," 29 Dec 2015. [Online]. Available: http://www.eia.gov/energy_in_brief/article/major_energy_sources_and_users.cfm. [Accessed 2 Dec 2016].

[13] NaturalGas.org, "Residential Uses," 20 Sep 2013. [Online]. Available: http://naturalgas.org/overview/uses-residential/. [Accessed 5 Nov 2016].

[14] J. Tobin, "Distribution of Natural Gas: The Final Step in the Transmission Process," Energy Information Administration, Office of Oil and Gas, June, 2008.

[15] NaturalGas.org, "Commercial Uses," 20 Sep 2013. [Online]. Available: http://naturalgas.org/overview/uses-commercial/. [Accessed 5 Nov 2016].

[16] NaturalGas.org, "Uses in Industry," 20 Sep 2013. [Online]. Available: http://naturalgas.org/overview/uses-industrial/. [Accessed 5 Nov 2016].

[17] U.S. Energy Information Administration, "Total U.S. Energy Production Increses for Sixth Consecutive Year," 29 March 2016. [Online]. Available: http://www.eia.gov/todayinenergy/detail.php?id=25852. [Accessed 20 June 2016].

[18] U.S. Energy Information Administration, "How the United States Uses Energy," U.S. Energy Information Administration, Apr 2016. [Online]. Available: http://www.eia.gov/energyexplained/?page=us_energy_use. [Accessed 6 Nov 2016].

[19] M. Phillips, "Must-Know: The Supply Chain Delivers Electricity," 5 Sep 2014. [Online]. Available: http://marketrealist.com/2014/09/must-know-supply-chain-delivers-electricity/. [Accessed 30 Nov 2016].

[20] HITACHI, "Smart Grid," HITACHI, 10 Mar 2015. [Online]. Available: http://www.hitachi.com/environment/showcase/solution/energy/smartgrid.html. [Accessed 30 Nov 2016].

[21] Newsletter for the European Union (NEU), "Public and Private Sectors Should Work Together for Smart Grids," 6 Jun 2016. [Online]. Available: http://www.newslettereuropean.eu/public-and-private-sectors-should-work-together-for-smart-grids/. [Accessed 30 Nov 2016].

[22] R. Williams, "Electric Transmission and Generation: How the Grid Works," Director of Business Development for AFL Global, 28 September 2011. [Online]. Available: https://www.youtube.com/watch?v=nJ-eBqEnraE. [Accessed 20 June 2016].

[23] U.S. Dept. of Energy, "TITLE XIII-Smart Grid Sec. 1301. Statement of Policy on Modernization of Electricity Grid," Dec 2017. [Online]. Available: http://energy.gov/sites/prod/files/oeprod/DocumentsandMedia/EISA_Title_XIII_Smart_Grid.pdf. [Accessed 7 Nov 2016].

[24] P. Smith, "FPL's 'Smart Grid' Enabled Quick Power Restoration After Hurricane Matthew," Miami Herald, 14 Oct 2016. [Online]. Available: http://www.miamiherald.com/opinion/op-ed/article108384897.html. [Accessed 7 Nov 2016].

[25] Office of Electricity Delivary and Energy Reliability, "What is the Smart Grid?," U.S. Department of Energy, [Online]. Available: https://www.smartgrid.gov/the_smart_grid/smart_grid.html. [Accessed 7 Nov 2016].

[26] Federal Energy Regulation Commision Staff Team, "Assessment of Demand and Advanced Metering," Federal Energy Regulation Commision, Dec, 2008.

[27] R. Williams, "Smart Grid Tutorial," Director of Business Development for AFL Global, 14 September 2011. [Online]. Available: https://www.youtube.com/watch?v=dD1vybH-uFI. [Accessed 20 June 2016].

[28] H. Hahn, S. Meyer-Nieberg and S. Pickl, "Electric Load Forecasting Methods: Tools for Decision Making," *European Journal of Operational Research,* vol. 199, no. 3, pp. 902-907, 2009.

[29] R. C. Smith, Uncertainty Quantification: Theory, Implementation, and Applications, Philadelphia: SIAM, 2014.

[30] National Research Council, Assessing the Reliability of Complex Models: Mathematical and Statistical Foundations of Verification, Validation, and Uncertainty Quantification, Washington, D.C.: The National Academies Press, 2012.

[31] T. Hong, "Energy Forecasting: Past , Present, and Future," *Foresight: The International Journal of Forecasting,* no. 32, pp. 43-48, Winter, 2014.

[32] T. Hong, "Crystal Ball Lessons in Predictive Analytics," *EnergyBiz,* pp. 35-37, 12-13 May 2015.

[33] D. W. Bunn and E. D. Farmer, "Economic and Operational Context of Electrical Load Prediction," in *Comparative Models for Electrical Load Forecasting*, New York, John Wiley and Sons, 1985, pp. 3-11.

[34] B. Hobbs, S. Jitprapaikulsarn, S. Konda, V. Chankong, K. A. Loparo and D. J. Maratukulam, "Analysis of the Value for Unit Commitment of Improved Load Forecasts," *IEEE Transaction on Power Systems,* vol. 14, no. 4, pp. 1342-1348, 1999.

[35] P. Pinson, Estimation of the Uncertainty in Wind Power Forecasting, Lyngby, Denmark: Ph.D. Dissertation, Dept. of Elec. Eng., Technical University of Denmark(DTU), 2006.

[36] H. Madsen, P. Pinson, P. Bacher, J. Kloppenborg, J. Tastu and E. B. Iversen, "Methods for Probabilistic Forecasting of Wind and Solar Power Generation," Department of Applied Mathematics and Computer Science and Center for IT-Intelligent Energy Systems (CITIES), Lyngby, Denmark, 2015.

[37] T. Hong and S. Fan, "Probabilistic Electricity Forecasting: A Tutorial Review," *International Journal of Forecasting,,* vol. 32, no. 3, pp. 1-32, 2014.

[38] G. E. Huck, A. A. Mahmoud, R. B. Comerford, J. Adams and E. Dawson, "Load Forecast Bibliography: Phase I," *IEEE Transactions on Power Apparatus and Systems,* Vols. PAS-99, no. 1, pp. 53-58, 1980.

[39] A. A. Mahmoud, T. H. Ortmeyer and R. E. Reardon, "Load Forecasting Bibliography Phase II," *IEEE Transactions on Power Apparatus and Systems,* Vols. PAS-100, no. 7, pp. 3217-3220, 1981.

[40] A. H. Murphy and R. L. Winkler, "Probability Forecasting in Meterology," *Journal of the American Statistical Association,* vol. 79, no. 387, pp. 489-500, 1984.

[41] I. Moghram and S. Rahman, "Analysis and Evaluation of Five Short-Term Load Forecasting Techniques," *IEEE Transaction on Power Systems,* vol. 4, no. 4, pp. 1484-1491, 1989.

[42] D. W. Bunn, "Forecasting Loads and Prices in Competitive Power Markets," *Proceedings of the IEEE,* vol. 88, no. 2, pp. 163-169, 2000.

[43] H. S. Hippert, C. E. Pedreira and R. C. Souza, "Neural Networks for Short-Term Load Forecasting: A Review and Evaluation," *IEEE Transactions on Power Systems,* vol. 16, no. 1, pp. 44-55, 2001.

[44] H. K. Alfares and M. Nazeeruddin, "Electric Load Forecasting: Literature Survey and Classification of Methods," *International Journal of Systems Science,* vol. 33, no. 1, pp. 23-34, 2001.

[45] T. Hong, Short Term Electric Load Forecasting, Raleigh, North Carolina: Ph.D. Dissertation, Dept. of Operational Research and Electrical Engineering, North Carolina State University (NCSU), 2010.

[46] S. Siddique, Automation of Energy Demand Forecasting, Milwaukee, Wisconsin, USA: Master's Thesis, Dept. of Elec. and Comp. Engg., Marquette University, 2013.

[47] T. Hong, P. Pinson and S. Fan, "Global Energy Forecasting Competition 2012," *International Journal of Forecasting,* vol. 30, no. 2, pp. 357-363, 2014.

[48] N. Amral, C. Ozveren and D. King, "Short Term Load Forecasting Using Multiple Linear Regression," in *Power Engineering Conference (UPEC)*, Brighton, U.K., 2007.

[49] T. Hong, P. Wang and H. L. Willis, "A Naïve Multiple Linear Regression Benchmark for Short Term Load Forecasting," in *IEEE Power and Energy Sosceity (PES) General Meeting*, Detroit, MI, 2011.

[50] P. Wang, B. Liu and T. Hong, "Electrical Load Forecasting With Recency Effect: A Big Data Approach," *International Journal of Forecasting,* vol. 32, no. 3, pp. 585-597, 2016.

[51] S. B. Taieb and R. J. Hyndman, "A Gradient Boosting Approach to the Kaggle Load Forecasting Competition," *International Journal of Forecasting,* vol. 30, no. 2, pp. 382-394, 2014.

[52] J. Xie, T. Hong and J. Stroud, "Long-Term Retail Energy Forecasting With Consideration of Residential Customer Attrition," *IEEE Transaction on Smart Grid,* vol. 6, no. 5, pp. 2245-2252, 2015.

[53] SAS Institute Inc., "The GLM Procedure," [Online]. Available: https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#glm_toc.htm. [Accessed 26 Dec 2016].

[54] T. Hong, P. Wang and L. White, "Weather Station Selection for Electric Load Forecasting," *International Journal of Forecasting,* vol. 31, no. 2, pp. 286-295, 2015.

[55] G. U. Yule, "On the Method of Investigating Periodicities in Disturbed Series, With Special Referece to Wolfer's Sunspot Numbers," *Philosophical Transactions of the Royal Society London,* vol. 226, no. Series A, pp. 267-298, 1927.

[56] D. N. Gujarati, Basic Econometrics, New York: The McGraw-Hill, 2004.

[57] G. Box and G. M. Jenkins, Time Series Analysis: Forecasting and Control, San Francisco: Prentice Hall, 1970.

[58] A. Jain and B. Satish, "Short Term Load Forecasting by Clustering Technique Based on Daily Average and Peak Loads," in *IEEE Power and Energy Society (PES) General Meeting*, Calgary, Alberta, 2009.

[59] M. Espinoza, J. A. K. Suykens, R. Belmans and B. D. Moor, "Electricity Load Forecasting," *IEEE Control Systems,* vol. 27, no. 5, pp. 43-57, 2007.

[60] S.-J. Huang and K.-R. Shih, "Short-Term Load Forecasting via ARMA Model Identification Including Non-Gaussian Process Considerations," *IEEE Transactions on Power Systems,* vol. 18, no. 2, pp. 673-679, 2003.

[61] S. Siddique and R. J. Povinelli, "Learning Energy Demand Domain Knowledge via Feature Transformation," in *IEEE Power and Energy Society (PES) General Meeting*, Washington, DC, 2014.

[62] R. Hyndman, A. B. Koehler, J. K. Ord and R. D. Snyder, Forecasting With Exponential Smoothing: The State Space Approach, Berlin, Germany: Springer Verlag, 2008.

[63] R. G. Brown, "Exponential Smoothing for Predicting Demand," Arthur D Little, Inc., Cambridge, Massachusetts, 1956.

[64] R. G. Brown, Smotting Forecasting and Prediction of Discrete Time Series, Englewood Cliffs, NJ : Prentice-Hall, 1963.

[65] C. C. Holt, "Forecasting Seasonals and Trends by Exponentially Weighted Averages," Office of Naval Research Memorandum 52, Carnegie Institute of Technology, 1957.

[66] P. R. Winters, "Forecasting Sales by Exponentially Weighted Moving Averages," *Management Science,* pp. 324-342, 1960.

[67] C. C. Holt, "Forecasting Trends and Seasonals by Exponentially Weighted Averages," *International Journal of Forecasting,* vol. 20, no. 1, pp. 5-10, 2004.

[68] J. G. D. Gooijer and R. J. Hyndman, "25 Years of Time Series Forecasting," *International Journal of Forecasting,* vol. 22, no. 3, pp. 443-473, 2006.

[69] J. W. Taylor and P. E. M. Lilian M. De Menezes, "A Comparison of Univariate Methods for Forecasting Electricity Demand up to a Day Ahead," *International Journal of Forecasting,* vol. 22, no. 1, pp. 1-16, 2006.

[70] M. Adya, F. Collopy, J. S. Armstrong and M. Kennedy, "Automatic Identification of Time Series Features for Rule-Based Forecasting," *International Journal of Forecasting,* vol. 17, no. 2, pp. 143-157, 2001.

[71] J.-j. Wang, D.-x. Niu and L. Li, "An ARMA Cooperate With Artificial Neural Network Approach in Short-Term Load Forecasting," in *International Conference on Natural Computation*, Tianjin, China, 2009.

[72] Z. H. Osman, M. L. Awad and T. K. Mahmoud, "Neural Network Based Approach for Short-Term Load Forecasting," in *IEEE Power Systems Conference and Exposition (PSCE)*, Seattle, WA, USA, 2009.

[73] P. Qingle and Z. Min, "Very Short-Term Load Forecasting Based on Neural Network and Rough Set," in *International Conference on Intelligent Computation Technology and Automation (ICICTA)*, Changsha, 2010.

[74] S. Ramos, J. Soares, Z. Vale and S. Ramos, "Short-Term Load Forecasting Based on Load Profiling," in *IEEE Power and Energy Society (PES) General Meeting*, Vancouver, BC, 2013.

[75] D. Xin-Hui, T. Feng and T. Shao-Qiong, "Study of Power System Short-Term Load Forecast Based on Artificial Neural Network and Genetic Algorithm," in *International Conference on Computational Aspects of Social Networks (CASoN)*, Taiyuan, China, 2010.

[76] X. Sun, P. B. Luh, L. D. Michel, S. Corbo, K. W. Cheung and W. Guan, "An Efficient Approach for Short-Term Substation Load Forecasting," in *IEEE Power and Energy Society (PES) General Meeting*, Vancouver, BC, 2013.

[77] S. Haykin, Kalman Fitering and Neural Networks, New York: John Wiley and Sons, 2001.

[78] A. Deoras, "Electricity Load and Price Forecasting Webinar Case Study," 10 Sep 2010. [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/28684-electricity-load-and-price-forecasting-webinar-case-study?s_cid=LF_OPTA_3. [Accessed 4 Oct 2016].

[79] "ISO New England," [Online]. Available: https://www.iso-ne.com/. [Accessed 26 Dec 2016].

[80] MathWorks Inc., "Electricity Load Forecasting Using Neural Networks," MATLAB, 10 Sep 2010. [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/28684-electricity-load-and-price-forecasting-webinar-case-study/content/Electricity%20Load%20&%20Price%20Forecasting/Load/html/LoadScriptNN.html. [Accessed 4 Oct 2016].

[81] MathWorks Inc., "Load Forecasting Using Bagged Regression Trees," 10 Sep 2010. [Online]. Available: https://www.mathworks.com/examples/matlabxl/community/19612-load-forecasting-using-bagged-regression-trees. [Accessed 4 Oct 2016].

[82] MathWorks Inc., "Electricity Price Forecasting With Neural Networks," 10 Sep 2010. [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/28684-electricity-load-and-price-forecasting-webinar-case-study/content/Electricity%20Load%20&%20Price%20Forecasting/Price/html/PriceScriptNN.html. [Accessed 4 Oct 2016].

[83] M. Mohri, A. Rostamizadeh and A. Talwalkar, Foundation of Machine Learning, Cambridge, Massachusetts: MIT Press, 2012.

[84] M. Mohandes, "Support Vector Machines for Short-Term Electrical Load Forecasting," *International Journal of Energy Research,* vol. 4, no. 335-345, p. 26, 2002.

[85] F. Shu and C. Luonan, "Short-Term Load Forecasting Based on an Adaptive Hybrid Method," *IEEE Transactions on Power Systems,* vol. 21, no. 1, pp. 392-401, 2006.

[86] "New York City Independent System Operator," [Online]. Available: http://www.nyiso.com/public/index.jsp. [Accessed 28 Dec 2016].

[87] X. Jin, J. Wu, Y. Dong and D. Chi, "Application of a Hybrid Model to Short-Term Load Forecasting," in *International Conference of Information Science and Management Engineering (ISME)*, Xi'an , 2010.

[88] A. Escobar M. and L. P. Perez, "Application of Support Vector Machines and ANFIS to the Short-Term Load Forecasting," in *IEEE Transmission and Distributioan Conference and Exposition*, Bogota, Colombia, 2008.

[89] Q. Ding, "Long-Term Load Forecast Using Decision Tree Method," in *IEEE Power Systems Conference and Exposition (PSCE)*, Atlanta, Georgia, USA, 2006.

[90] J. R. Quinlan, "Induction of Decision Trees," *Machine Learning,* vol. 1, no. 1, pp. 81-106, 1986.

[91] M. B. A. Hamid and T. K. A. Rahman, "Short Term Load Forecasting Using an Artificial Neural Network Trained by Artificial Immune System Learning Algorithm," in *International Conference on Computer Modelling and Simulation*, Cambridge, United Kingdom, 2010.

[92] L. N. de Castro and F. J. V. Zuben, "Artificial Immune System : Part I – Basic Theory and Applications," Technical Report, RT - DCA 01/99, Dec, 1999.

[93] L. N. de Castro and F. J. V. Zuben, "Learning and Optimisation Using the Clonal Selection Principle," *IEEE Transactions on Evolutionary Computation,* vol. 6, no. 3, pp. 239-251, 2002.

[94] Y. Y. Hsu and K. L. Ho, "Fuzzy Expert Systems: An Application to Short-Term Load Forecasting," *IEEE Proceedings - Generation, Transmission and Distribution,* vol. 139, no. 6, pp. 471-477, 1992.

[95] D. K. Ranaweera, N. F. Hubele and G. G. Karady, "Fuzzy Logic for Short Term Load Forecasting," *Electrical Power and Energy Systems,* vol. 18, no. 4, pp. 215-222, 1996.

[96] S. C. Pandian, K. Duraiswamy, C. C. A. Rajan and N. Kanagaraj, "Fuzzy Approach for Short Term Load Forecasting," *Electric Power Systems Research,* vol. 76, no. 6-7, pp. 541-548, 2006.

[97] S. Ahmadi, H. Bevrani and H. Jannaty, "A Fuzzy Inference Model for Short-Term Load Forecasting," in *Iranian Conference on Renewable Energy and Distributed Generation*, Tehran, Iran, 2012.

[98] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *Annals of Statistics,* vol. 29, no. 5, pp. 1189-1232, 2001.

[99] D. Mease and A. Wyner, "Evidence Contrary to the Statistical View of Boosting : A Rejoinder to Responses," *Journal of Machine Learning Research,* vol. 9, no. 1, pp. 131-156, 2008.

[100] T. J. Hastie and R. Tibshirani, Generalized Additive Models, London: Chapman and Hill, 1995.

[101] R. J. Hyndman and S. Fan, "Monash Electricity Forecasting Model," Business and Economic Forecasting Unit, Australia, 28 May 2014.

[102] W. E. Cooke, "Forecasts and Verifications in Western Australia," *Monthly Weather Review,* vol. 34, pp. 23-24, 1906.

[103] L. A. Hughes, "Probability Forecasting: Reasons, Procedures, Problems," NOAA Technical Memorandum, Silver Spring, 1980.

[104] C. Chatfield, "Calculating Interval Forecasts," *Journal of Business and Economic Statistics,* vol. 11, no. 2, pp. 121-135, 1993.

[105] C. Chatfield, "Prediction Intervals," *Principles of Forecasting,* vol. 30, no. May, pp. 475-494, 2001.

[106] C. Chatfield, "A Survey of Recent Developments in Forecasting Methods," *Journal of Chemical Information and Modeling,* vol. 53, no. 9, pp. 1689-1699, 2013.

[107] A. E. Raftery, "Use and Communication of Probabilistic Forecasts," University of Washington, Washington, 2014.

[108] D. Spiegelhalter, M. Pearson and I. Short, "Visualizing Uncertainty About the Future," *Science,* vol. 333, no. 6048, pp. 1393-1400, 2011.

[109] Y. Zhang, J. Wang and X. Wang, "Review on Probabilistic Forecasting of Wind Power Generation," *Renewable and Sustainable Energy Reviews,* vol. 32, no. C, pp. 255-270, 2014.

[110] T. Gneiting, F. Balabdaoui and A. E. Raftery, "Probabilistic Forecasts, Calibration and Sharpness," *Journal of the Royal Statistical Society,* vol. 69, no. 2, pp. 243-268, 2007.

[111] T. Gneiting and M. Katzfuss, "Probabilistic Forecasting," *Annual Review of Statistics and its Application,* vol. 1, pp. 125-151, 2014.

[112] T. Hong, P. Pinson, S. Fan, H. Zareipour, A. Troccoli and R. J. Hyndman, "Probabilistic Energy Forecasting: Global Energy Forecasting Competetion 2014 and Beyond," *International Journal of Forecasting,* vol. 32, pp. 896-913, 2016.

[113] T. Gneiting and A. E. Raftery, "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association,* vol. 102, no. 477, pp. 359-378, 2007.

[114] R. L. Winkler, "A Decision-Theoretic Approach to Interval Estimation," *Journal of the American Statistical Association,* vol. 67, no. 337, pp. 187-191, 1972.

[115] P. Pinson, "Wind Energy: Forecasting Challenges for its Operational Management," *Statistical Science,* vol. 28, no. 4, pp. 564-585, 2013.

[116] Bank of England, "Inflation Report Fan Charts," Aug 2015. [Online]. Available: http://www.bankofengland.co.uk/publications/Documents/inflationreport/2015/augfc.pdf. [Accessed 7 Jan 2017].

[117] H. Raiffa and R. Schlaifer, Applied Statistical Decision Theory, Boston: Division of Research, Harvard Business School, 1961.

[118] P. Shah, "Newsvendor Problem (part 1)," YouTube, 21 April 2014. [Online]. Available: https://www.youtube.com/watch?v=6BUF_jxU5p8. [Accessed 20 June 2016].

[119] P. Shah, "Newsvendor Problem (part 2)," YouTube, 21 April 2014. [Online]. Available: https://www.youtube.com/watch?v=8o_7K0vOdUg. [Accessed 20 June 2016].

[120] T. Gneiting, "Quantiles as Optimal Point Forecasts," *International Journal of Forecasting,* vol. 27, no. 2, pp. 197-207, 2011.

[121] R. Weron, "Electricity Price Forecasting: A Review of the State-of-the-art With a Look Into the Future," *International Journal of Forecasting,* vol. 30, no. 4, pp. 1030-1081, 2014.

[122] NORD POOL, "Historical Market Data," [Online]. Available: http://www.nordpoolspot.com/historical-market-data/. [Accessed 7 Jan 2017].

[123] B. Rossi, *Density Forecasts in Economics, Forecasting and Policymaking,* Barcelona: ICREA-Universitat Pompeu Fabra, 2015.

[124] F. X. Diebold, T. A. Gunther and A. S. Tay, "Evaluating Density Forecasts With Applications to Financial Risk Management," *International Economic Review,* vol. 39, no. 4, pp. 863-883, 1998.

[125] P. Gaillard and Y. Goude, "Additive Models and Robust Aggregation for GEFCom2014 Probabilistic Electric Load and Electricity Price Forecasting," *International Journal of Forecasting,* vol. 32, no. 3, pp. 1-13, 2015.

[126] Puget Sound Energy, "Our History," Puget Sound Energy, [Online]. Available: http://pse.com/aboutpse/CorporateInfo/Pages/Our-History.aspx. [Accessed 10 Jan 2017].

[127] P. Engle, C. Granger, R. Ramanathan and F. Vahid-Araghi, "Probabilistic Methods in Forecasting Hourly Loads," Quantitive Economic Research Inc. (QUERI), San Diego, California, 1993.

[128] T. Hong, J. Wilson and J. Xie, "Long Term Probabilistic Load Forecasting and Normalization With Hourly Information," *IEEE Transactions on Smart Grid,* vol. 5, no. 1, pp. 456-462, 2014.

[129] R. Littell, W. Stroup and R. Freund, SAS for Linear Models, Cary, NC: SAS Institute Inc., 2002.

[130] J. Xie, T. Hong, T. Laing and C. Kang, "On Normality Assumption in Residual Simulation for Probabilistic Load Forecasting," *IEEE Transactions on Smart Grid,* vol. PP, no. 99, pp. 1-8, 2015.

[131] J. Xie and T. Hong, "GEFCom2014 Probabilistic Electric Load Forecasting: An Integrated Solution With Forecast Combination and Residual Simulation," *International Journal of Forecasting,* vol. 32, no. 3, pp. 1012-1016, 2016.

[132] A. Pierrot and Y. Goude, "Short-Term Electricity Load Forecasting With Generalized Additive Models," in *International Conference on Intelligent System Application to Power Systems (ISAP)*, Hersonissos, Greece, 2011.

[133] T. Hastie and R. Tibshirani, Generalized Additive Models, London: Chapman and Hall, 1990.

[134] S. N. Wood, Generalized Additive Models: An Introduction With R, London: Chapman and Hall, 2006.

[135] S. Vitullo, "Error Evaluation on Unusual Days," in *Disaggregating Time Series Data for Energy Consumption by Aggregate and Individual Customer*, Milwaukee, WI, USA, Ph.D. Dissertation, Dept. of Elec. and Comp. Engg., Marquette University, 2011, pp. 103-107.

[136] S. N. Wood, S. Shaw and Y. Goude, "Generalized Additive Models for Large Data Sets," *Journal of the Royal Statistical Society: Series C (Applied Statistics),* vol. 64, no. 1, pp. 139-155, 2015.

[137] S. Fan and R. Hyndman, "Short-Term Load Forecasting Based on Semi-Parametric Additive Model," *IEEE Transaction on Power Systems,* vol. 27, no. 1, pp. 134-141, 2012.

[138] R. Hyndman and S. Fan, "Density Forecasting for Long-Term Peak Electricity Demand," *IEEE Transactions on Power Systems,* vol. 25, no. 2, pp. 1142-1153, 2010.

[139] V. Dordonnat, A. Pichavant and A. Pierrot, "GEFCom2014 Probabilistic Electric Load Forecasting Using Time Series and Semi-Parametric Regression Models," *International Journal of Forecasting,* vol. 32, no. 3, pp. 1005-1011, 2016.

[140] R. T. Clemen, "Combining Forecasts: A Review and Annotated Bibliography," *International Journal of Forecasting,* vol. 5, no. 4, pp. 559-583, 1989.

[141] C. Genest and J. V. Zidek, "Combining Probability Distribution: A Critique and an Annotated Bibliography," *Statistical Sciences,* vol. 1, no. 1, pp. 114-148, 1986.

[142] P. Goodwin, "New Evidence on the Value of Combining Forecasts," *Foresight: International Journal of Applied Forecasting,* vol. 35, no. 12, pp. 33-36, Winter 2009.

[143] J. Armstrong, Combining Forecasts, 1-19: Kluwer Academic Publishers, 2001.

[144] G. Cheng and Y. Yang, "Forecast Combination With Outlier Protection," *International Journal of Forecasting,* vol. 31, no. 2, pp. 223-237, 2015.

[145] R. W. Koenker and G. Bassett, "Regression Quantiles," *Econometrica,* vol. 46, no. 1, pp. 33-50, 1978.

[146] R. W. Koenker and K. F. Hallock, "Quantile Regression: An Introduction," *Symposium on Econometric Tools,* vol. 15, no. 4, pp. 1-24, 2000.

[147] R. W. Koenker and K. F. Hallock, "Quantile Regression," *Journal of Economic Perspectives,* vol. 15, no. 4, pp. 143-156, 2001.

[148] B. S. Cade and B. R. Noon, "A Gentle Introduction to Quantile Regression for Ecologists," *Frontiers in Ecology and the Environment,* vol. 1, no. 8, pp. 412-420, 2003.

[149] A. Katchova, "Quantile Regression," Econometrics Academy, 24 Feb 2013. [Online]. Available: https://sites.google.com/site/econometricsacademy/econometrics-models/quantile-regression. [Accessed 17 Jan 2016].

[150] X. He and Y. Wei, "Tutorial on Quantile Regression," Eastern North American Region (ENAR), 2005. [Online]. Available: http://avesbiodiv.mncn.csic.es/estadistica/curso2011/qr9.pdf. [Accessed 17 Jan 2016].

[151] J. Nowotarski and R. Weron, "Computing Electricity Spot Price Prediction Intervals Using Quantile Regression and Forecast Averaging," *Computational Statistics,* vol. 30, no. 3, pp. 791-803, 2015.

[152] B. Liu, J. Nowotarski and T. Hong, "Probabilistic Load Forecasting via Quantile Regression Averaging on Sister Forecasts," in *IEEE Transactions on Smart Grid*, 2015.

[153] I. Steinwart and A. Christmann, "Estimating Conditional Quantiles With the Help of the Pinball Loss," *Bernoulli,* vol. 17, no. 1, pp. 211-225, 2011.

[154] K. Maciejowska, J. Nowotarski and R. Weron, "Probabilistic Forecasting of Electricity Spot Prices Using Factor Quantile Regression Averaging," *International Journal of Forecasting,* vol. 32, no. 3, pp. 957-965, 2016.

[155] K. Maciejowska and J. Nowotarski, "GEFCom2014 Competition and Probabilistic Electricity Price Forecasting," *International Journal of Forecasting,* vol. 32, no. 1, pp. 1051-1056, 2016.

[156] P. Gaillard, Y. Goude and R. Nedellec, "Additive Models and Robust Aggregation for GEFCom2014 Probabilistic Electric Load and Electricity Price Forecasting," *International Journal of Forecasting,* vol. 32, no. 3, pp. 1-13, 2016.

[157] S. Haben and G. Giasemidis, "A Hybrid Model of Kernel Density Estimation and Quantile Regression (Load Forecasting)," *International Journal of Forecasting,* vol. 32, no. 3, pp. 1017-1022, 2016.

[158] J. Jeon and J. W. Taylor, "Short-Term Density Forecasting of Wave Energy Using ARMA-GARCH Models and Kernel Density Estimation," *International Journal of Forecasting,* vol. 32, no. 3, pp. 991-1004, 2016.

[159] "FINO1 Research Platform," [Online]. Available: http://www.fino1.de/. [Accessed 15 Jan 2017].

[160] E. Mangalova and O. Shesterneva, "Sequence of Nonparametric Models for GEFCom2014 Probabilistic Electric Load Forecasting," *International Journal of Forecasting,* vol. 32, no. 3, pp. 1023-1028, 2016.

[161] R. Kharkar, "Lasso Regression," 20 Aug 2015. [Online]. Available: https://www.youtube.com/watch?v=jbwSCwoT51M. [Accessed 15 Jan 2017].

[162] F. Ziel and B. Liu, "Lasso Estimation for GEFCom2014 Probabilistic Electric Load Forecasting," *International Journal of Forecasting,* vol. 32, no. 3, pp. 1029-1037, 2016.

[163] S. B. Taieb, R. Huser, R. J. Hyndman and M. G. Genton, "Forecasting Uncertainty in Electricity Smart Meter Data by Boosting Additive Quantile Regression," in *IEEE Transactions on Smart Grid*, 2016.

[164] M. Landry, T. P. Erlinger, D. Patschke and C. Varrichio, "Probabilistic Gradient Boosting Machines for GEFCom2014 Wind Forecasting," *International Journal of Forecasting,* vol. 32, no. 3, pp. 1061-1066, 2016.

[165] G. I. Nagy, G. Barta, S. Kazi, G. Borbély and G. Simon, "GEFCom 2014: Probabilistic Solar and Wind Power Forecasting Using a Generalized Additive Tree Ensemble Approach," *International Journal of Forecasting,* vol. 32, no. 3, pp. 1087-1093, 2016.

[166] R. Juban, H. Ohlsson, M. Maasoumy, L. Poirier and J. Z. Kolter, "A Multiple Quantile Regression Approach to the Wind, Solar, and Price Tracks of GEFCom2014," *International Journal of Forecasting,* vol. 32, no. 3, pp. 1094-1102, 2016.

[167] H. Quan, D. Srinivasan and A. Khosravi, "Uncertainty Handling Using Neural Network-Based Prediction Intervals for Electrical Load Forecasting," *Energy,* vol. 73, no. 14, pp. 916-925, 2014.

[168] G. Dudek, "Multilayer Perceptron for GEFCom2014 Probabilistic Electricity Price Forecasting," *International Journal of Forecasting,* vol. 32, no. 3, pp. 1057-1060, 2016.

[169] Y. Zhang and J. Wang, "GEFCom2014 Probabilistic Solar Power Forecasting Based on K-Nearest Neighbor and Kernel Density Estimator," in *IEEE Power and Energy Society (PES) General Meeting*, Denver, CO, 2015.

[170] Y. Zhang and J. Wang, "K-Nearest Neighbors and a Kernel Density Estimator for GEFCom2014 Probabilistic Wind Power Forecasting," *International Journal of Forecasting,* vol. 32, no. 3, pp. 1074-1080, 2016.

[171] E. Mangalova and O. Shesterneva, "K-Nearest Neighbors for GEFCom2014 Probabilistic Wind Power Forecasting," *International Journal of Forecasting,* vol. 32, no. 3, pp. 1067-1073, 2016.

[172] J. Huang and M. Perry, "A Semi-Empirical Approach Using Gradient Boosting and K-Nearest Neighbors Regression for GEFCom2014 Probabilistic Solar Power Forecasting," *International Journal of Forecasting,* vol. 32, no. 3, pp. 1081-1086, 2016.

[173] R. J. Hyndman and A. B. Koehler, "Another Look at Measures of Forecast Accuracy," *International Journal of Forecasting,* vol. 22, no. 4, p. 679–688, 2006.

[174] P. E. Tetlock and D. Gardner, Superforecasting: The Art and Science of Prediction, New York: Crown, 2015.

[175] P. Pinson, J. K. M. Henrik Aa. Nielsen, H. Madsen and G. N. Kariniotakis, "Non-Parametric Probabilistic Forecasts of Wind Power: Required Properties and Evaluation," *Wind Energy,* vol. 10, no. 6, pp. 497-516, 2007.

[176] G. W. Brier, "Verification of Forecasts Expressed in Terms of Probaility," *Monthly Weather Review,* vol. 78, no. 1, pp. 1-3, 1950.

[177] A. H. Murphy, "A New Vector Partition of the Probability Score," *Journal of Applied Meteorology,* vol. 12, no. 4, pp. 595-600, 1973.

[178] A. H. Murphy, "Skill Scores Based on the Mean Square Error and Their Relationships to the Correlation Coefficient," *Monthly Weather Review,* vol. 116, no. 12, pp. 2417-2424, 1988.

[179] A. H. Murphy, "Hedging and Skill Scores for Probability Forecasts," *Journal of Applied Meteorology,* vol. 12, no. 1, pp. 215-223, 1973.

[180] A. H. Murphy, "A Sample Skill Score for Probability Forecasts," *Monthly Weather Review,* vol. 102, no. 1, pp. 48-55, 1974.

[181] J. Hernández-Orallo, P. Flach and C. Ferri, "Brier Curves: A New Cost-Based Visualisation of Classifier Performance," in *International Conference on Machine Learning (ICML)*, Bellevue, Washington, 2011.

[182] J. Hernández-Orallo, P. Flach and C. Ferri, "Loss, A Unified View of Performance Metrics: Translating Threshold Choice Into Expected Classification," *Journal of Machine Learning Research,* vol. 13, pp. 2813-2869, 2012.

[183] A. H. Murphy, "The Ranked Probability Score and the Probability Score: A Comparison," *Monthly Weather Review,* vol. 98, no. 12, pp. 917-924, 1970.

[184] B. Rossi and T. Sekhposyan, "Evaluating Predictive Densities of US Output Growth and Inflation in a Large Macroeconomic Data Set," *International Journal of Forecasting,* vol. 30, no. 3, pp. 662-682, 2014.

[185] B. Rossi and T. Sekhposyan, "Alternative Tests for Correct Specification of Conditional Predictive Densities," Working Paper, Spain, 2016.

[186] B. Rossi and T. Sekhposyan, "Conditional Predictive Density Evaluation in the Presence of Instabilities," *Journal of Economics,* vol. 177, no. 2, pp. 199-212, 2013.

[187] R. N. Allan, A. M. L. da Silva and R. C. Burchett, "Evaluation Methods and Accuracy in Probabilistic Load Flow Solutions," *IEEE Transactions on Power Apparatus and Systems,* Vols. PAS-100, no. 5, pp. 2539-2546, 1981.

[188] N. L. Johnson, "Systems of Frequency Curves Generated by Methods of Translation," *Biometrika,* vol. 36, no. 1/2, pp. 149-176, 1949.

[189] R. Povinelli and M. Saber, *Probabilistic Natural Gas Demand Forecasting Using Kernel Density Estimator,* Milwaukee, WI: Unpublished, 2016.

[190] MathWorks Inc., "Normal Distribution," [Online]. Available: https://www.mathworks.com/help/stats/normal-distribution.html. [Accessed 16 Feb 2017].

[191] F. Pukelsheim, "The Three Sigma Rule," *The American Statistician,* vol. 48, no. 2, pp. 88-91, 1994.

[192] R. L. Ott and M. Longnecker, An Introduction to Statistical Methods and Data Analysis, Boston, USA: Brooks Cole; 7th ed., 2015.

[193] N. A. Weiss, Introductory Statistics, Essex, England: Pearson; 10th ed., 2015.

[194] C. H. Brase, Understanding Basic Statistics, Boston, USA: Brooks Cole; 7th ed., 2015.

[195] N. L. Johnson, "Bivariate Distributions Based on Simple Translation Systems," *Biometrika,* vol. 36, no. 3/4, pp. 297-304, 1949.

[196] MathWorks Inc., "Kernel Distribution," [Online]. Available: https://www.mathworks.com/help/stats/kernel-distribution.html. [Accessed 16 Feb 2017].

[197] A. D'Silva, "Kernel Density Estimation," in *Estimating the Extreme Low-Temperature Event Using Nonparametric Methods*, Milwaukee, Wisconsin, USA, Mater's Thesis, Dept. of Elec. and Comp. Engg., Marquette University, 2014, pp. 41-44.

[198] A. W. Bowman and A. Adelchi, Applied Smoothing Techniques for Data Analysis, New York: Oxford University Press, 1997.

[199] D. L. Jones, "Johnson Curve Toolbox for MATLAB: Analysis of Non-Normal Data Using the Johnson Family of Distributions," MathWorks Inc., College of Marine Science, University of South Florida, St. Petersburg, Florida, USA, 2014.

[200] UNISTAT Statistical Software, "Data Transformation," UNISTAT Ltd., [Online]. Available: https://www.unistat.com/guide/quality-control-data-transformation/. [Accessed 8 Mar 2017].

[201] J. Draper, "Properties of Distributions Resulting From Certain Simple Transformations of the Normal Distribution," *Biometrika,* vol. 39, p. 290–301, 1952.

[202] R. E. Wheeler, "Quantile Estimators of Johnson Curve Parameters," *Biometrika,* vol. 67, pp. 725-728, 1980.

[203] I. D. Hill, R. Hill and R. L. Holder, "Algorithm AS 99: Fitting Johnson Curves by Moments," *Journal of the Royal Statistical Society: Series C (Applied Statistics),* vol. 25, no. 2, pp. 180-189, 1976.

[204] I. D. Hill, "Algorithm AS 100: Normal-Johnson and Johnson-Normal Transformations," *Journal of the Royal Statistical Society. Series C (Applied Statistics),* vol. 25, no. 2, pp. 190-192, 1976.

[205] G. Box and D. Cox, "An Analysis of Transformations," *Journal of the Royal Statistical Society: Series B (Methodological),* vol. 26, no. 2, pp. 211-252, 1964.

[206] Minitab Inc., "Data Transformations for Capability Analysis," [Online]. Available: http://support.minitab.com/en-us/minitab/17/topic-library/quality-tools/capability-analyses/distributions-and-transformations-for-nonnormal-data/data-transformations/. [Accessed 22 Jan 2017].

[207] R. H. Brown, S. R. Vitullo, G. F. Corliss, M. Adya, P. E. Kaefer and R. J. Povinelli, "Detrending Daily Natural Gas Consumption Series to Improve Short-Term Forecasts," in *IEEE Power and Energy Society (PES) General Meeting*, Denver, CO, July, 2015.

[208] R. H. Brown, D. Kaftan, G. F. Corliss and R. J. Povinelli, "Improving Natural Gas Forecasting by Combining Models," in *International Symposium on Forecasting*, Cairns, Australia, June, 2017.

[209] T. Gao, Blending as a Multi-Horizon Time Series Forecasting Tool, Milwaukee, Wisconsin, USA: Master's Thesis, Dept. of Elec. and Comp. Engg., Marquette University, May, 2014.

[210] J. Z. Amedu, Probabilistic Age Grading From Near Infrared Spectroscopy, Bagamoyo, Tanzania: Master's Thesis, African Institute for Mathematical Sciences (AIMS), June, 2017.

[211] J. Z. Amedu, M. P. Milali, M. Saber, M. Sikulu-Lord and G. F. Corliss, "Probabilistic Mosquito Age Grading From Near Infrared Spectroscopy," *in preperation,* 2017.