Design Day Analysis - Forecasting Extreme Daily Natural Gas Demand

David Joseph Kaftan
Marquette University

DESIGN DAY ANALYSIS - FORECASTING EXTREME DAILY NATURAL GAS DEMAND

by

David Kaftan, B.S.

A Thesis Submitted to the Faculty of the
Graduate School, Marquette University,
in Partial Fulfillment of the Requirements for
the Degree of Master of Science

Milwaukee, Wisconsin

May 2018

ABSTRACT

DESIGN DAY ANALYSIS - FORECASTING EXTREME DAILY NATURAL GAS DEMAND

David Kaftan, B.S.

Marquette University, 2018

This work provides a framework for Design Day analysis. First, we estimate the temperature conditions which are expected to be colder than all but one day in N years. This temperature is known as the Design Day condition. Then, we forecast an upper bound on natural gas demand when temperature is at the Design Day condition.

Natural gas distribution companies (LDCs) need to meet demand during extreme cold days. Just as bridge builders design for a nominal load, natural gas distribution companies need to design for a nominal temperature. This nominal temperature is the Design Day condition. The Design Day condition is the temperature that is expected to be colder than every day except one in N years. Once Design Day conditions are estimated, LDCs need to prepare for the Design Day demand. We provide an upper bound on Design Day demand to ensure LDCs will be able to meet demand.

Design Day conditions are determined in a variety of ways. First, we fit a kernel density function to surrogate temperatures - this method is referred to as the Surrogate Kernel Density Fit. Second, we apply Extreme Value Theory - a field dedicated to finding the maxima or minima of a distribution. In particular, we apply Block-Maxima and Peak-Over-Threshold (POT) techniques. The upper bound of Design Day demand is determined using a modified version of quantile regression.

Similar Design Day conditions are estimated by both the Surrogate Kernel Density Fit and Peaks-Over-Threshold methods. Both methods perform well. The theory supporting the POT method and the empirical performance of the SKDF method lends confidence in the Design Day conditions estimates. The upper bound of demand on these conditions is well modeled by the modified quantile regression technique.

ACKNOWLEDGMENTS

David Kaftan, B.S.

First and foremost, I could not have performed this research without the aid of my committee. Each of them served me as invaluable mentors. Drs. Povinelli, Brown, and Corliss' passion for education is unparalleled by every professor I have had thus far.

Secondly, I must thank the GasDay Lab - not just for financially sponsoring this work, but also for creating a community of scholars I have the privilege of calling my friends. Without the camaraderie of graduate students - Paul, Andrew Kirkham, Andrew Tran, Greg, Masabho, Jarrett, Anisha, Saber, Michele, and Heidi - this thesis would have driven me insane. The undergraduates who aided me are regrettably too many to list here, though I feel the need to thank everyone on the MATLAB Development team - including Marielle, Jeffrey, Tom, Colleen, Shivani, and Chandan. Their efforts allowed me to focus more heavily on research.

Finally, I need to thank my family, Isha, and Birdhouse for their incredible emotional support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction to Design Day Analysis

This work provides a framework for Design Day analysis. First, we estimate the temperature conditions which are expected to be colder than all but one day in N years. This temperature is known as the Design Day condition. Then, we forecast an upper bound on natural gas demand when temperature is at the Design Day condition.

This chapter provides an introduction to the problem. First, we describe the natural gas industry. We also describe the GasDay Lab, from which the work in this thesis originates. Finally, we introduce broadly the problem addressed in this thesis: planning for the Design Day.

1.1 The United States Natural Gas Industry

Natural gas typically refers to methane gas (CH_4) - a fossil fuel by product of biological decomposition. Much of the natural gas in the United States is produced from shale plays, or sedimentary rock with natural gas trapped in its pores. Natural gas has become the largest source of energy in the United States [12]. The large

increase in U.S. natural gas production can be attributed to technological advancements in refining shale.

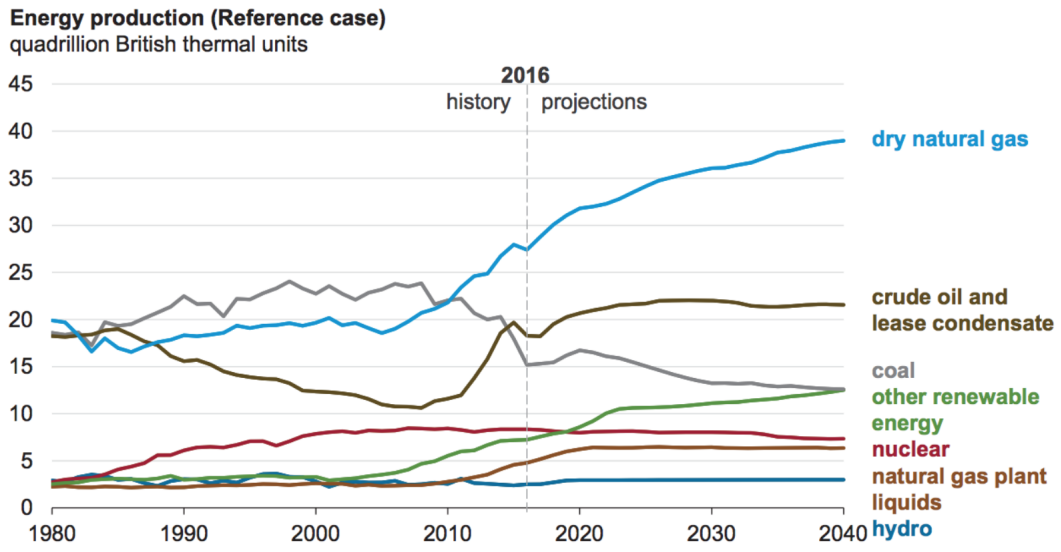


Figure 1.1: Energy production in United States [43]

Natural gas production has greatly increased in recent years. In 2010, it overtook coal as the largest source of energy in the United States [43].

The U.S. Energy Information Administration categorizes the uses of natural gas into five cases: electric power, industrial, residential, commercial, transportation. Local Distribution Companies (LDCs) are responsible for delivering natural gas to industrial, residential, and commercial end users. Depending on state regulations, the end users may buy gas directly from the LDC or from marketers. LDCs are still responsible for delivering gas bought by end users from marketers.

It takes time for LDCs to bring gas onto their system. LDCs purchase gas in

units of dekatherms (Dth) - a unit of energy. They must forecast future gas demand on their system to supply it on time. An error in forecasting leads to errors in supply. If too much gas is planned, LDCs may face a penalty for leaving gas in the pipeline. If too little gas is planned, LDCs may have to buy expensive gas on the spot market, or - in the extreme case - run out. Therefore, good natural gas forecasting plays a vital role in the economics and safety of the natural gas infrastructure.

1.2 Marquette University GasDay™

Marquette University GasDay™ is a research lab and small business that develops tools to forecast natural gas for LDCs. Our flagship product forecasts daily natural gas demand with an eight day horizon. 37 LDCs across the United States have signed license agreements with Marquette University GasDay™, providing over one million days of historical natural gas demand data. Our business has expanded to include analysis of extreme cold scenarios, known as Design Days.

1.3 Forecasting Natural Gas Demand

To plan for meeting natural gas demand during extreme cold events, we must forecast natural gas demand. Demand is forecast for regions of customers known as *operating areas*. Several methods are used for forecasting natural gas demand. Two

of the most common modeling techniques are *linear regression* and *artificial neural networks*. These two modeling techniques are common ways to map inputs - such as daily temperature - to predicted demand [32, 44].

The inputs that are used in natural gas forecasting fall into three categories: weather, calendar, and autoregressive.

1.3.1 Weather Effects on Gas Demand

Natural gas is used largely for heating space. If the temperature is cold, the demand increases. Temperature is the largest factor driving demand for most of the LDCs that GasDay supports. Due to the non-linear relationship between temperature and demand, we transform temperature. The most common transformation is the Heating Degree Day (HDD). This is calculated accordingly

$$\text{HDD} = \max(\textit{Reference} - \textit{Temperature}, 0), \tag{1.1}$$

providing a piece-wise linear relationship. The *Reference* is the temperature below which the relationship between demand and HDD is approximately linear. Typical values for *Reference* include 65 (HDD65) and 55 (HDD55). In Figure 1.2, HDD55 would fit the data much better than HDD65. Alternatively, we could use both HDD55 and HDD65 as inputs.

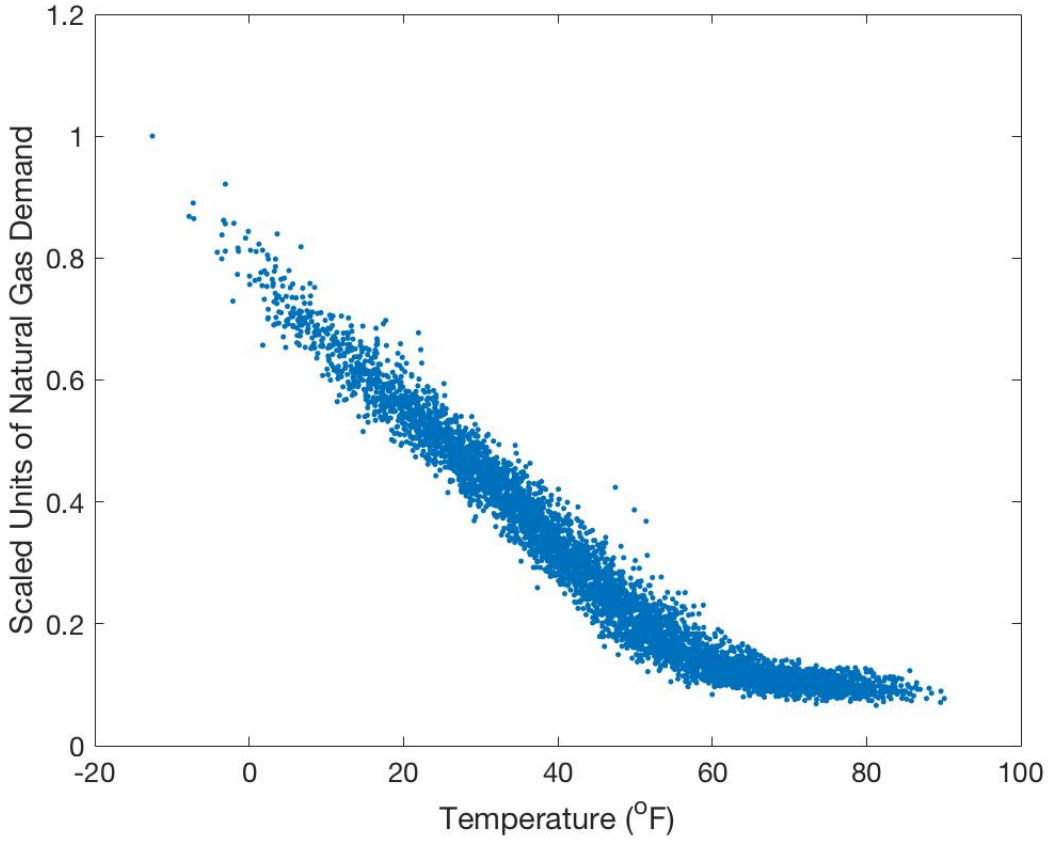


Figure 1.2: Natural gas demand vs temperature at midwestern utility

Natural gas demand increases during cold weather. HDD55 captures the approximately linear relationship between demand and temperature when temperature is less than 55 ^oF.

While temperature on the day of forecasting is the primary driver of demand, temperature on the previous day and wind also affect demand. In order to incorporate wind into linear models, we adjust the HDD accordingly [44]:

$$\text{HDDW} = \begin{cases} \frac{\text{Wind Speed}+152}{160} \times \text{HDD}, \text{ Wind Speed} \leq 8 \\ \\ \frac{\text{Wind Speed}+72}{80} \times \text{HDD}, \text{ Wind Speed} > 8. \end{cases} \tag{1.2}$$

.

To address the affect of the prior day temperature, Kaefer introduces a metric known as Prior Day Weather Sensitivity (PDWS) [24]. PDWS is the ratio of coefficients $-\frac{\beta_2}{\beta_1}$ in the linear regression model $flow = \beta_0 + \beta_1 HDD + \beta_2 \Delta HDD$, where ΔHDD is the change in temperature from yesterday. Ishola calculates PDWS at different temperatures. Ishola shows that PDWS is smaller in magnitude as the temperature gets colder [41]. Figure 1.3 shows the raw PDWS at different temperatures along with an exponential fit.

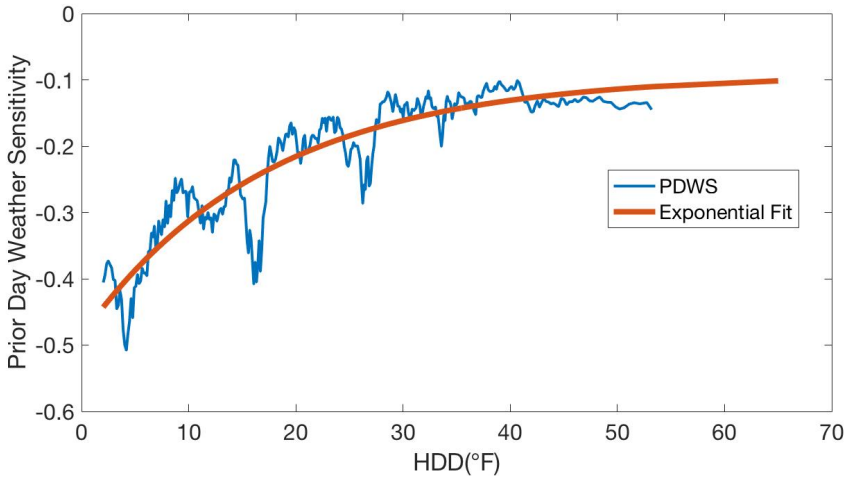


Figure 1.3: Prior Day Weather Sensitivity

Ishola fit this linear regression model to different subsets of data. The subsets of data are created by sliding a window. He fit an exponential decay to the relationship between temperature and PDWS. This provides evidence that there is a temperature dependency to PDWS.

Of course, there are drivers of demand outside of weather. These may impact the demand during an extreme cold event, so we discuss them below.

1.3.2 Non-Weather Effects on Gas Demand

Natural gas demand is highly dependent on the calendar. In particular, the day of week affects the gas demand. More gas is used during the week due commercial and industrial use [44]. The gas demand on the previous day has also been shown to be a good predictor of gas demand on the current day [44].

GasDay employs these forecasting techniques to assist LDCs in preparing for the Design Day.

1.4 The Design Day

Just as bridge builders design their bridges to withstand a nominal load, LDCs design their systems to withstand a nominal day of gas demand. **The Design Day** is a hypothetical day invented by an LDC to characterize an extreme case of daily natural gas demand. LDCs use the Design Day as a benchmark on which to base their long term plans. The Design Day typically is characterized by the temperature that could drive an extreme amount of natural gas demand, known as the Design Day conditions.

1.4.1 Estimating Design Day Conditions

Design Day conditions are the weather conditions (typically temperature) that are expected to be more extreme than all but one day in the next N years. For this reason, the Design Day conditions also are known as the One-in-N conditions. LDCs face the challenge of estimating the Design Day conditions. Consider estimating the One-in-40 condition. The definition of the One-in-40 condition can vary from LDC to LDC. For some LDCs, the One-in-40 temperature is expected to be exceeded $\frac{1}{40}$ times this year. Some LDCs are concerned only with the coldest day of the year; they want the temperature that is expected to be exceeded *at least* once in a year with a $\frac{1}{40}$ probability. These two definitions correspond to Block-Maxima and Peaks Over Threshold approaches in extreme value theory, respectively. Estimating temperatures in the extreme tail of the temperature distribution is the main challenge of estimating Design Day conditions. Once the Design Day conditions are estimated, we can estimate the Design Day demand.

1.4.2 Estimating Design Day Demand

Design Day demand is the demand expected to occur on a day with Design Day conditions. We cannot be exactly sure what the flow will be on a day with Design Day conditions. For this reason, it is important to quantify the uncertainty in Design Day demand. For example, we might forecast 100 MDth of

gas demand on a day with Design Day conditions. It would be more helpful to say that we are 99% certain demand will be less than 110 MDth. The practitioner could prepare for 110 MDth to mitigate the risk of being under-prepared for the Design Day.

1.4.3 Quantifying Performance of Design Day Analysis

Its important to analyze the performance of our Design Day analysis in a way that reflects its usefulness to GasDay's customers. The predicted Design Day conditions should be exceeded once every N years. The forecast of Design Day demand should accurately reflect the uncertainty of demand given Design Day conditions.

1.5 Problem Statement

The problem addressed in this thesis is threefold. First, we estimate Design Day conditions. Second, we quantify the uncertainty in demand on the Design Day. Finally, we quantify the usefulness of our analysis to our customers.

1.6 Thesis Roadmap

Following this introduction, Chapter 2 provides the background of Design Day analysis. Chapter 2 provides the resources and tools that are applied and extended in Chapter 3. Chapter 3 proposes methods for determining Design Day conditions, Design Day demand, and evaluating their performance. These methods are experimented on in Chapter 4. Chapter 5 summarizes the thesis and outlines future work.

CHAPTER 2

Design Day Analysis State-of-the-Art

The purpose of this chapter is to discuss background required to understand this thesis, state-of-the-art of Design Day analysis, and related research. The first section discusses the state-of-the-art methods for determining Design Day conditions. The second section discusses the state of the art methods for determining Design Day demand. The third and final section discusses the state of the art methods for evaluating the performance of Design Day conditions and Design Day demand forecasts.

2.1 Determining the Design Day Conditions

Determining the Design Day conditions is the first step in Design Day analysis. Design Day conditions can be determined in many ways. In this section, we will cover the current practices in industry and Extreme Value Theory - a field that looks at the statistics of unlikely events.

2.1.1 Current Practice in Industry

Practitioners typically determine the Design Day conditions one of two ways. Some LDC's choose the coldest day in the last N years (N typically ranges from 10 to 50). Other LDC's fit a distribution to temperatures and choose the temperature with the cumulative density function equal to $1/N$ [10]. GasDay has come across LDCs who use more advanced techniques. One anonymous LDC described using the Gumbel Distribution - a method that will be described in 2.1.2. Many LDCs rely on GasDay to provide Design Day conditions. GasDay has developed its own method for determining the Design Day condition - Surrogate Data Kernel Density Fit.

Surrogate Data Kernel Density Fit

The Surrogate Data Kernel Density Fit (SKDF) is described fully in [25]. Its process is broken down into two steps. First, temperature data is augmented with the use of surrogate data. Second, a kernel density function is fit to the data.

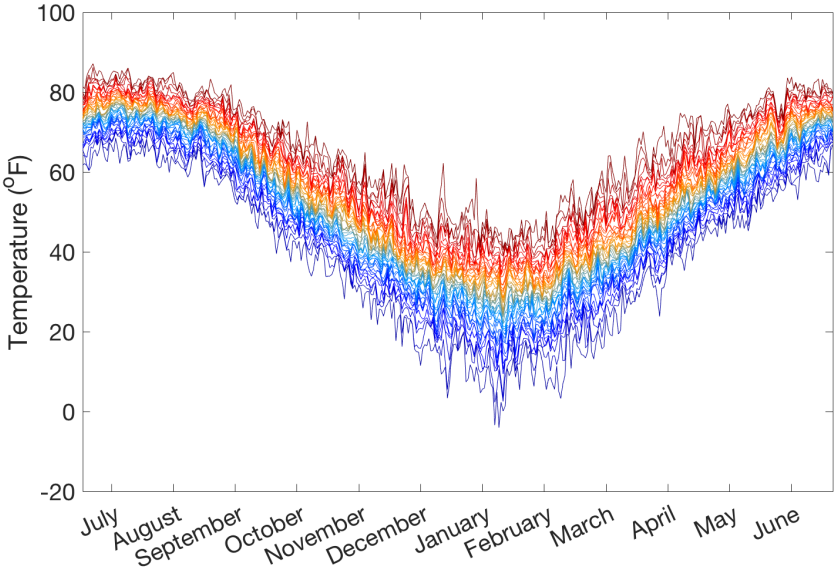
The temperature data augmentation increases the amount of data in the cold tail of the temperature distribution using surrogate data. Consider every temperature that has ever been recorded at a particular weather station on January 1st. The SKDF estimates what those temperatures would have looked like on January 2nd. First, January 1st temperatures shifted by the difference between the

mean January 2nd temperature and the mean January 1st temperature. Next, the January 1st temperatures are scaled by the ratio of lower standard deviations of January 2nd to 1st. The transformed January 1st temperatures now have the same mean and standard deviation as January 2nd. If the set of transformed temperatures is merged with the set of temperatures on January 2nd, there are now twice as much data on January 2nd. We call these new temperatures surrogate data because they were transformed from a surrogate source. This can be repeated by transforming the temperatures that occurred on December 31 or January 3rd. GasDay does this for the closest 90 calendar days - resulting in 91 times the original amount of data for January 2nd. This entire process is repeated for every day in winter, resulting in 91 times the original number of winters in the dataset. The result of the surrogate data method is visualized in Figure 2.1. D'Silva derived a method for determining the coldest 91 days of the year [11]. Using D'Silva's method instead of the calendar winter season ensures that the coldest day of each year will be included in the set of cold temperatures.

A kernel density function is fit to all surrogate temperatures in the winter. The One-in-N temperature can be determined by finding the temperature at which the cumulative density function equals $\frac{1}{N*NumberofDaysInWinter}$.

The SKDF is shown to perform well in practice [25]. For a more theoretical perspective, we turn to Extreme Value Theory.

(A)



(B)

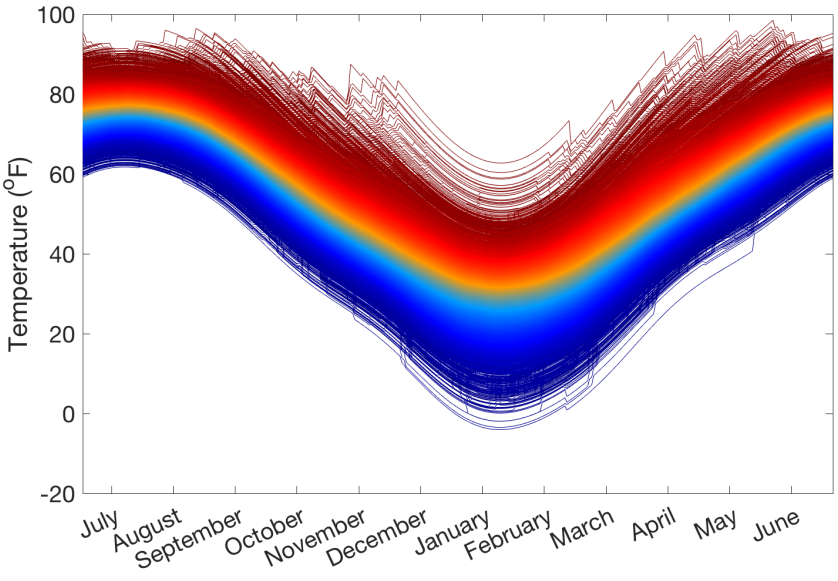


Figure 2.1: Surrogate and empirical temperature Quantiles

The daily average temperatures for each day of year are sorted and plotted according to their rank. The coldest temperature for each day of year appears in dark blue, while the hottest appears in dark red. (A) shows 45 years of raw temperature data, while (B) shows the dataset after the surrogate data transformation. Clearly, the surrogate data provides a more coherent picture of temperature distributions across the year.

2.1.2 Extreme Value Theory

Scientists have been looking into the extreme tails of distributions for centuries. As early as 1709 Nicholas Bernoulli determined the greatest expected duration of life for a group of men [30]. Since then, Extreme Value Theory has birthed two classes of methods: Block-Maxima and Peaks over Threshold.

Block-Maxima

The Block-Maxima approach to extreme values focuses on modeling the maximum (or minimum) value of a sampled distribution. For example, a normal distribution is sampled 100 times, and the samples are ordered minimum to maximum. This process is repeated to get several sets of 100 samples, each ordered minimum to maximum. The maximum sample will vary each time 100 samples are taken. The Block-Maxima approach attempts to model the distribution of maximum samples.

In practice, the name "Block-Maxima" is much more intuitive. Consider a 50 years of temperature data. The 50 years are broken up into 1-year *blocks*. The maximum of each block is taken, and the distribution of maxima is modeled. The characteristics of this distribution are the foundations of extreme value theory.

The field of extreme values was first laid out in 1928 when R.A. Fisher and

L.H.C. Tippet developed three distributions to which the block maxima taken from different distributions converge [16]. These include the Gumbel, Fréchet [13], and Weibull. Their cumulative distribution function is

$$F(x) = \begin{cases} \exp(-(1+\xi(x-\mu)/\sigma)^{(-1/\xi)}), & \text{if } \xi \neq 0 \\ \exp(\exp(-(x-\mu)/\sigma)), \text{ if } \xi = 0 \end{cases} \tag{2.1}$$

where μ is the mean, σ is the scale, and ξ is the shape. When $\xi > 0$, it is a Type II distribution - also known as the Fréchet distribution. When $\xi < 0$, it is a Type III distribution - the mirror of a Weibull distribution. When $\xi = 0$, it is a Type I distribution - the mirror of a Gumbel distribution. Together, Equation (2.1) is known as the Generalized Extreme Value (GEV) distribution.

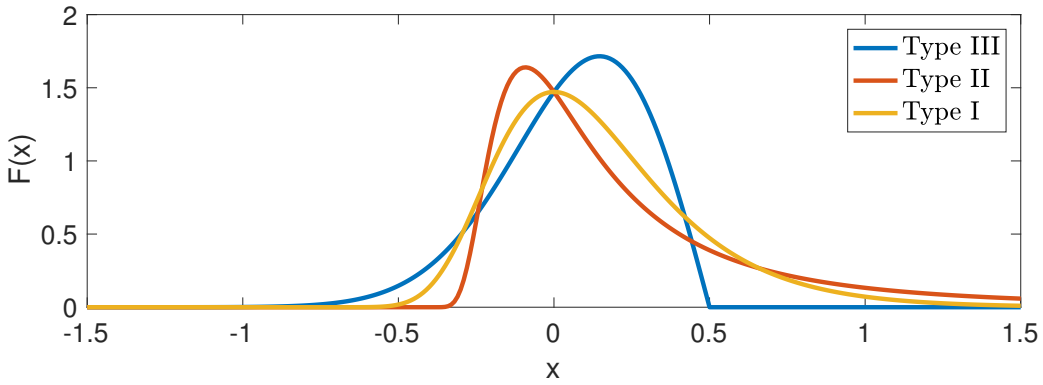


Figure 2.2: Generalized Extreme Value probability density functions

Applying the GEV distribution is straightforward for temperatures. As

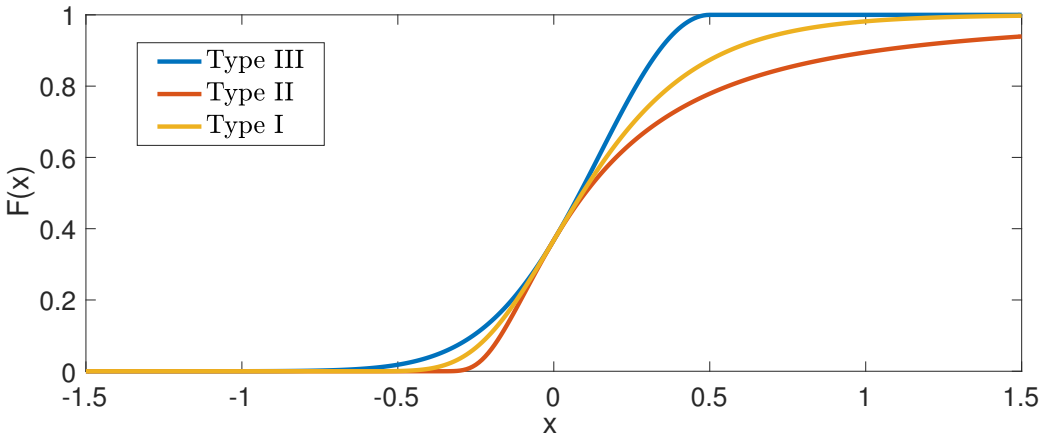


Figure 2.3: Generalized Extreme Value cumulative density functions

mentioned before, temperatures are split up into annual blocks. This is typical practice for weather extremes [2]. The GEV distribution is fit to the set of yearly maxima. The GEV distributions model the *maximum* of temperature distributions - for Design Day conditions, we are interested in the *minimum*. This can be remedied by simply multiplying our temperature data by -1.

The history of the Block-Maxima approach makes it a very appealing choice for determining Design Day conditions. It was both developed for and applied to determining the statistics of extreme weather events; estimating Design Day conditions requires modeling the statistics of extreme cold temperatures. Gumbel developed Block-Maxima approaches in his analysis of floods (1941, 1944, 1945, 1949) [30]. More recently, Hasan et al. used GEV distributions to characterize annual maximum temperatures [17].

Peaks Over Threshold (POT)

The implementation of the Peaks Over Threshold follows its name; set a threshold that is expected to be exceeded infrequently and fit a distribution to the data that exceeds the threshold. This is fundamentally modeling something very different from the Block-Maxima method. While the Block-Maxima method models only the coldest temperature of each year, the POT method is independent of time; it models the coldest temperatures no matter when they occurred. The idea of modeling the peaks over a threshold is described in detail by Davison and Smith [9]. The peaks over a threshold are often fit to a Generalized Pareto distribution, whose cumulative density function is

$$F(x) = \begin{cases} 1 - (1 + \xi(x - \mu)/\sigma)^{-1/\xi}, & \text{if } x \neq 0 \\ 1 - exp(-(x - \mu)/\sigma), & \text{if } x = 0. \end{cases} \tag{2.2}$$

Much of the motivation for using the Generalized Pareto distribution comes from analysis of the GEV distribution. Consider the following scenario for random process X. If the times between exceedances over the threshold follow a Poisson distribution, and the exceedances follow a Pareto distribution, then the Block-Maxima of X follows a Generalized Extreme Value distribution [9].

In addition to its relationship to extreme value distributions, there are

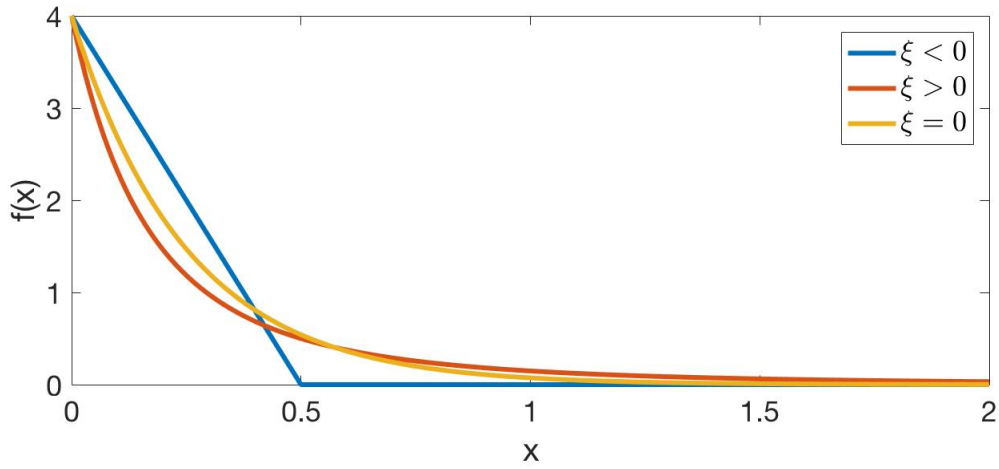


Figure 2.4: Probability density function for Generalized Pareto distribution

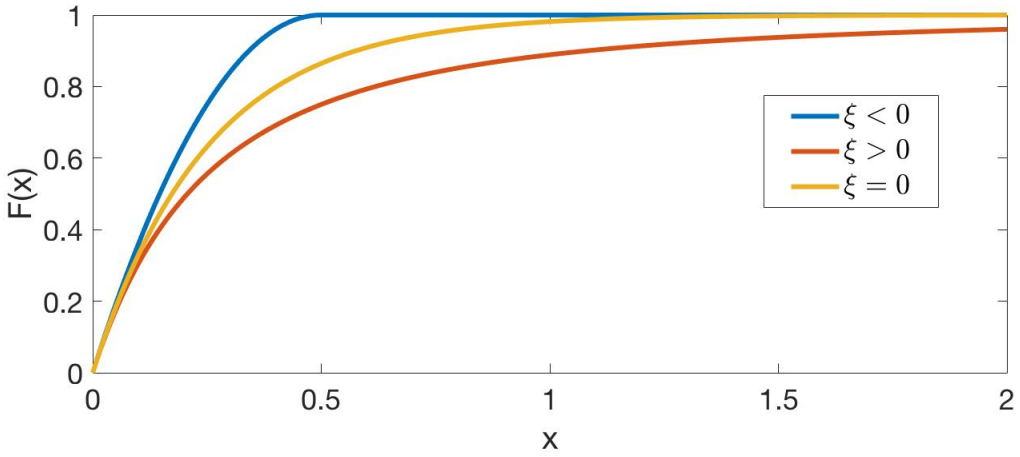


Figure 2.5: Cumulative density function for Generalized Pareto distribution

several useful properties of the Generalized Pareto distribution that hold given a high enough threshold. However, if the threshold is too high, there will not be enough data on which to fit the distribution. Therefore, one of the engineering challenges for using the Generalized Pareto distribution - and POT methods in general - is to determine what threshold to set. Typically, the threshold is set

analytically. A tool for analyzing where the threshold belongs arises from a property of the Generalized Pareto distribution; the average exceedence of the threshold should be linear with the threshold. Choosing a linear region in a plot of mean exceedence vs. threshold produces a valid threshold [2]. It may not be feasible to determine visually a valid threshold; it prevents the POT method from being automated. In this case, the threshold is set to some empirical quantile [39]. For example, we might choose to fit the Generalized Pareto distribution to the coldest two percent of days for each dataset.

Similar to the Block-Maxima method, the history of the POT method makes it appealing for determining Design Day conditions. Much of the POT research is motivated by characterizing extreme climate. Determining Design Day conditions requires modeling the distribution of extreme cold temperatures. Davison and Smith apply POT methods to model the distributions of extreme river flows and wave heights [9]. Gong applied the Generalized Pareto distribution to identify extreme temperature events [15].

We now have the background required to estimate Design Day conditions. We next look at the background required to determine Design Day demand based on these conditions.

2.2 Determining the Design Day Demand

Finding the demand associated with the Design Day conditions is an essential part of preparing for the Design Day. At the end of the day, LDCs will need to be prepared to meet demand on a Design Day. To assist LDCs in meeting demand, we explore two fields of literature: forecasting during extreme cold events and probabilistic forecasting.

2.2.1 Gas Forecasting During Extreme Cold Events

Understanding the relationship between demand and extreme cold weather is the primary challenge of forecasting the Design Day demand. Extreme cold days are the most important days to have accurate gas forecasts. Because more gas is used when it is extremely cold, more is at stake if a forecast performs poorly. For this reason, much research has been dedicated to forecasting during cold conditions. For example, Broehl used a linear regression model to estimated the natural gas demand given extreme cold weather [3]. Brown et al. built regression models from monthly data to forecast Design Day demand [4].

The primary challenge of forecasting energy during extreme cold events is overcoming the inherent data sparsity. Extreme cold events happen rarely, so there is always relatively less demand data on which to build a model. Kaefer addresses

this problem directly by transforming surrogate data from multiple operating areas to behave like a single operating area [24]. The resulting dataset has more data on the extreme cold days.

2.2.2 Probabilistic Forecasting

There will always be some amount of error when forecasting demand, especially during extreme cold days. LDCs are interested in the probability that the actual demand is much higher than the forecast - particularly during the Design Day; it is safer to be over-prepared for the Design Day than under-prepared. Instead of providing a *single estimate* (known as a *point forecast*) for the Design Day demand, we can provide a *CDF of possible demand* given the Design Day conditions (known as a *probabilistic forecast*). The usefulness of a probabilistic forecast is illustrated in Figure 2.6. Both plots (left and right) show simulated data that is unrealistically easy to forecast. The plot on the left shows a point forecast with respect to HDD. The point forecast is sufficient for determining the expected demand for some HDD. However, if a practitioner is more interested in the demand that has a 10% chance of being exceeded, they need to use the probabilistic forecast (right).

Probabilistic forecasting has been a popular topic in energy research. Saber created probabilistic forecast of hourly gas demand by fitting a distribution to the

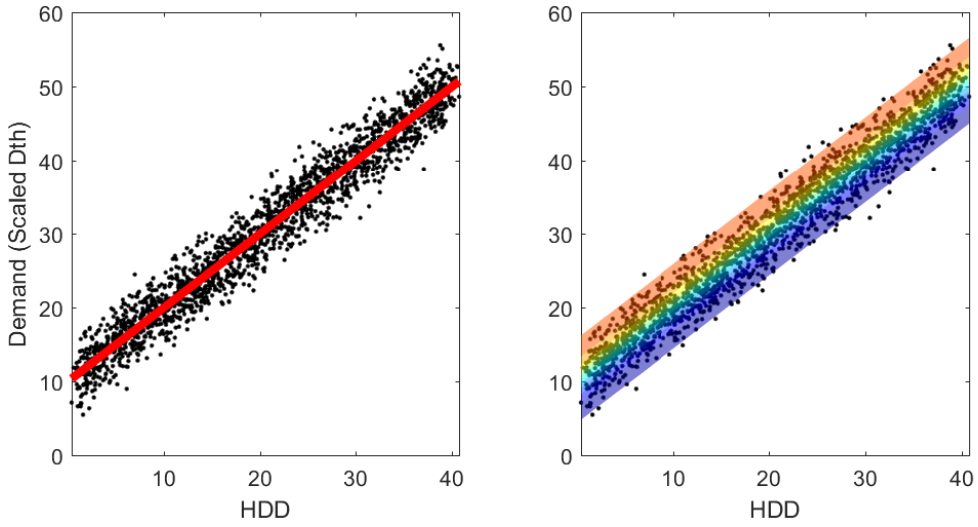


Figure 2.6: Point Forecast vs. Probabilistic Forecast

This figures above contain synthetic demand and HDD data. The left shows the least-squares regression line superimposed on the data. The right shows different probability bands from the quantile regression. Each band contains approximately 10% of the points. The plot on the right shows there is a 10% chance the demand will exceed 54 Dth given HDD= 40.

validation error of a forecast model [38]. Hong et al. [19] create probabilistic forecasts of monthly peak electric loads. Hong also provides a tutorial review of probabilistic forecasting in [20]. In particular, there is a large tutorial on probabilistic short term load forecasting - a section that is important to Design Day analysis as it focuses on forecasting single day demand. One of the main methods discussed is quantile regression.

Quantile Regression

Quantile regression was first introduced by Koenker and Bassett [27] as a method to find a linear relationship between the probability distribution of a target variable (i.e., flow) and some input variable (i.e., temperature). In other words, it is a method for deriving a conditional cumulative density function. The *quantile* is simply another word for the probability the variable of interest will be less than a threshold. For random variables X and Y, consider the equation $Y = X\theta$. To find the n^{th} quantile, θ satisfies

$$P(Y < X\theta) = \frac{n}{100}. \tag{2.3}$$

θ can be estimated via quantile regression. Quantile regression estimates θ by minimizing the Pinball Loss Score -

$$\text{Pinball Loss}(x, y, \theta, q) = \frac{2}{N} \sum_{i=0}^{N-1} \begin{cases} q|x_i\theta - y|, & \text{for } (x_i\theta - y) < 0 \\ (1-q)|x_i\theta - y|, & \text{for } (x_i\theta - y) > 0, \end{cases} \tag{2.4}$$

where q is the quantile of interest divided by 100, N is the number of samples in the set, x is the input matrix, and y is the target variable. Note that the Pinball Loss is equivalent to the mean absolute error when $q = 0.5$.

There are several methods for minimizing the Pinball Loss Score, starting

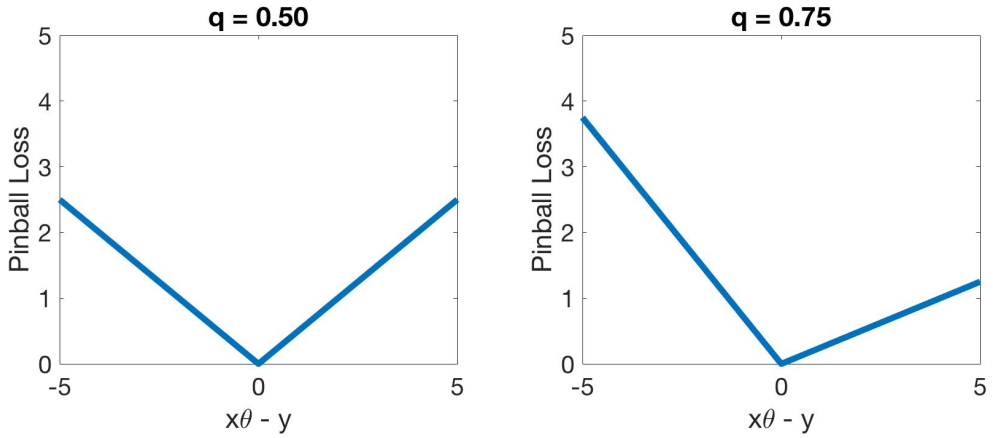


Figure 2.7: Pinball Loss Function for Two Different Quantiles

The Pinball Loss function for the 50^{th} quantile (left) is equivalent to mean absolute error. On the right is the Pinball Loss for the 75^{th} quantile. Notice that there is a larger penalty when $x_i\theta < y$. This is intuitive, because the quantile should fit above 75% of all y.

with gradient descent. Because the gradient is undefined when $x_i\theta - y = 0$, Zheng proposes smoothing the Pinball Loss function around $x_i\theta - y = 0$ such that the gradient is continuous. Zheng found that gradient descent on a smoothed Pinball Loss had higher accuracy than other gradient descent methods [48]. Hunter developed a majorize-minimize method for determining θ [23]. The most popular method is formulating quantile regression as a linear programming problem [29]. Quantile regression can then be solved by the simplex method [29] or by an interior point method [28] - two common linear programming techniques.

Quantile regression is unstable at extremely high quantiles. Modifications to quantile regression have been made to stabilize the extreme high quantiles. Wang et al. suggests using composite quantile regression [46], which constrains several

quantile regression models to have the same slope; the bias term distinguishes the different quantiles. For example, consider modeling demand with respect to HDD. If we want the model for the 99th quantile, we estimate the models for the 99, 89, and 79th quantiles. We constrain the HDD coefficients to be the same for each of the models, but allow the bias terms to differ. The 99th quantile is represented by the shared HDD coefficient and its unique bias term.

Composite quantile regression minimizes the weighted sum of the Pinball Loss function across different quantiles. This method was first introduced by Hogg in personal communication with Koenker [18]. The method is described in detail in [26]. Wang originally uniformly weighted the loss of each quantile, but later found that optimally weighting the Pinball Loss function for each quantile leads to an improvement in efficiency [47].

We have now discussed sufficient background to forecast Design Day conditions and demand. We now discuss the desired properties of these forecasts.

2.3 Forecasting Performance Metrics

Once Design Day conditions and demand have been forecast, their usefulness needs to be quantified. In this section, we discuss what properties of the forecasts matter to the practitioners: metrics used to quantify performance of Design Day

condition forecasts, and metrics used to quantify performance of the probabilistic forecasts.

2.3.1 What Matters to the Practitioner?

Practitioners want Design Day conditions to be exceeded - on average - once every N years. Practitioners are also concerned when Design Day conditions change each year they are calculated. Design Day conditions are often used to plan more than one year in the future [1]. If Design Day conditions that are forecast in 2016 differ greatly from the conditions forecast in 2017, LDCs will have to greatly change their long term plans. Therefore, it is important that Design Day condition forecasts do not change much over time.

Practitioners want point forecasts of Design Day demand to be as accurate as possible. They also want the probabilistic forecast to accurately reflect the certainty of the forecast on the Design Day. For example, the 99^{th} quantile should have a 1% chance of being exceeded on the Design Day.

2.3.2 Design Day Condition Metrics

D'Silva proposes a method to evaluate the performance of Design Day conditions [25]. The method consists of evaluating the number of times a Design Day condition is exceeded, then comparing that with the number of expected

exceedances of the condition. The actual number of exceedances is divided by the expected number of exceedances and called the actual-vs-expected ratio. The closer the actual-vs-expected ratio is to 1, the better the performance.

To address the need for Design Day conditions to not change drastically over time, we use the volatility metric Seaman introduced for retail sales forecasting [40]:

$$\text{volatility} = \sum_{iYear=0}^{nYears-1} (\text{Forecast}_{iYear} - \text{Forecast}_{iYear+1})^2. \tag{2.5}$$

Ideally, the volatility of a forecast is low, meaning that it does not change much over time. Taking the squared difference year-to-year makes sense in the context of a changing climate. While a changing climate might cause a trend in the Design Day condition forecast, we would expect the change to be slow over time. We would not expect huge peaks or valleys year to year, which taking the squared difference amplifies.

2.3.3 Probabilistic Forecasting Metrics

There are three concepts that determine the performance of a probabilistic forecast. The first concept - *reliability* - refers to a quantile having the correct number of data points above and below it given a large data set [36]. Reliability is sometimes referred to as calibration. *Sharpness* refers to the width of the confidence

bounds. For example, it is more useful to have a forecast that says there is a 90% chance that the demand between 4.99Dth and 5.01Dth than between 4Dth and 6Dth. Tighter probability bounds yield more useful forecasts. *Resolution* refers to the changing in probabilities based on the feature variables in a model. Consider the simplest probabilistic demand forecast - fitting a distribution to all demand. The forecasting method could have good reliability, but it tells us nothing about how the distributions ought to change with temperature.

The difficulty in evaluating the performance of a probabilistic forecast is that we do not know the ground truth of our forecast; the true probability distribution of demand for a day is unknown. There are several methods for evaluating performance of probabilistic forecasts. The most straightforward way to evaluate performance is to use the Pinball Loss - the loss on which quantile regression is optimized. The Pinball Loss function does not ensure reliability. Reliability is often considered to be the most important measure of probabilistic forecasting; it is to this effect that many papers suggest improving sharpness dependent on keeping the model reliable [34]. This implies a need to quantify reliability.

For each probabilistic forecast, there is only one observation of demand. How then, can we determine reliability, which is characterized by having the correct number of observations between quantiles? In practice, this problem is solved using the Probability Integral Transform (PIT). The PIT classifies a demand observation

by the quantiles of the probabilistic forecast it falls between. For example, if the observed demand is between the 70^{th} and 80^{th} forecast quantiles, it is placed into the $0.7 \leq F(x) < 0.8$ bin. This transformation is made for every forecast, resulting in a set of observed classes. We know the number of points that should be a part of each class (i.e., 10% of the points should be between the 70^{th} and 80^{th} quantiles). Most methods for quantifying reliability involve examining the PIT diagram shown in Figure 2.8 [14]. Transforming to a uniform distribution allows us to aggregate every observation of demand - each of which comes from a different distribution - into a single distribution. If the aggregated data is near uniform distributed, then our predicted distribution is close to the correct distribution, as shown in Figure 2.8. Saber suggests taking the mean absolute difference of a modified PIT diagram and the nominal percent of points in each bin [38]; this metric is known as the *percentage quantile calibration score* (PQCS).

We now have a background in the state-of-the-art methods for determining Design Day conditions and Design Day demand. Next, we develop the background into Design Day analysis.

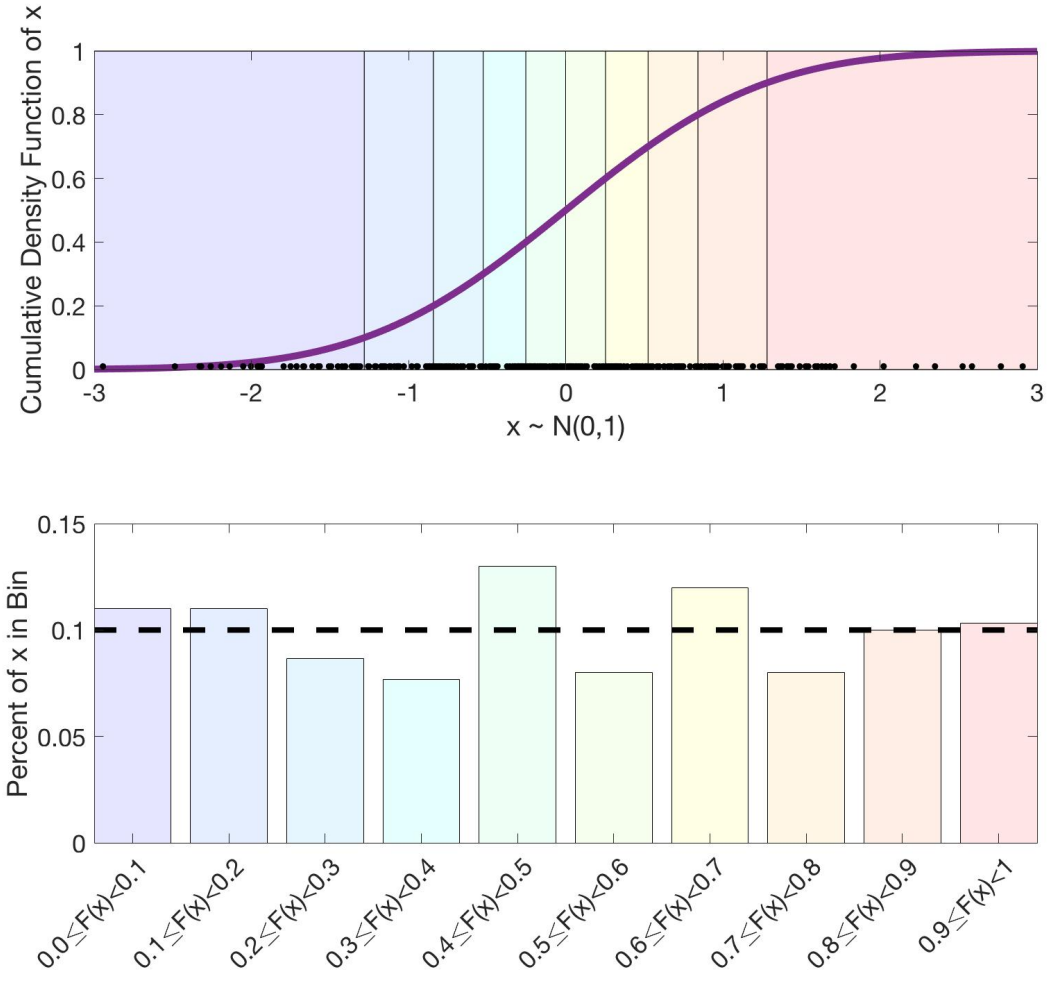


Figure 2.8: Probability Integral Transformation Diagram

(Above) The cumulative density function of X is displayed in purple. The background colors represent the 10% of the distribution. The black dots along the x-axis represent samples of X. (Below) The samples of X are input to the cumulative density function $F(x)$. The output of that function for all samples is made into the histogram. Given enough samples, the histogram approaches a uniform distribution (with height of the dotted black line) if $F(x)$ is the true cumulative density function of X.

CHAPTER 3

Methods for Design Day Analysis

In this chapter, we discuss our methods for Design Day analysis. First, we work through methods for determining the Design Day conditions. Next, we develop methods for forecasting the Design Day demand and the uncertainty in that forecast. Finally, we develop methods for evaluating the quality of our Design Day conditions and Design Day demand forecast.

3.1 Forecasting Design Day Conditions

Estimating Design Day conditions is split into two steps. First, we adjust temperature to make it a better predictor of natural gas demand. Then, we fit distributions to the adjusted temperature. Finally, we evaluate the model for the One-in-N condition.

3.1.1 Adjusting Temperature

As mentioned in Section 1.3.1, temperature is not the only influence of natural gas demand. Wind and the temperature from the previous day also make an impact. Rather than separately model the One-in-N condition for wind or previous

day, we adjust temperature. By adjusting temperature, we can incorporate wind and prior day temperature into our One-in-N condition without needing to model a multivariate distribution. Using an adjusted temperature, we can get a better predictor of natural gas demand than with temperature alone.

Wind is one of the primary predictors of natural gas demand. The equation for wind adjusted temperature - derived from Equations (1.1) and (1.2) - is

$$\text{Wind Adjusted Temp} = \begin{cases} 65 - (65 - \text{Temp})\frac{\text{Wind Speed}+152}{160}, \text{Wind Speed} \leq 8 \\ 65 - (65 - \text{Temp})\frac{\text{Wind Speed}+72}{80}, \text{Wind Speed} > 8. \end{cases} \tag{3.1}$$

Similar to HDDW, a temperature combined with a wind speed less than 8 miles per hour results in a warmer temperature. A wind speed greater than 8 miles per hour results in a colder temperature, as seen in Figure 3.1. We therefore expect more gas to be demanded on days with high wind.

To adjust temperature by the prior day effect, we first calculate the Prior

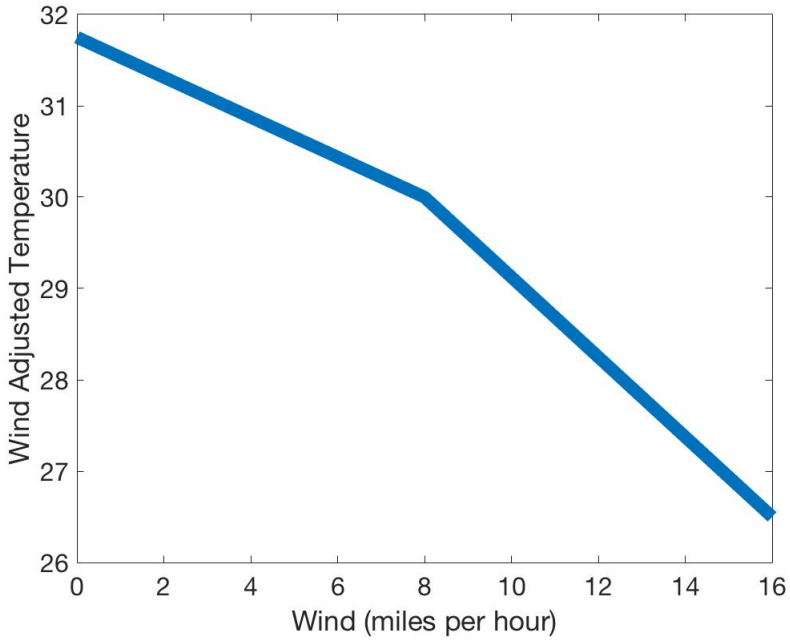


Figure 3.1: Wind adjustment for temperature = 30

Given a temperature of 30 degrees F, the wind adjusted temperature across a range on wind values is display above. A wind less than 8 mph adjusts the temperature warmer. A wind greater than 8 mph adjusts the temperature colder. A wind greater than 8 mph has a greater impact on the adjusted temperature, hence the steeper slope when wind $>$ 8 mph.

Day Weather Sensitivity (PDWS) according to the method described in Section

1.3.1. For each day k, we calculate the prior day adjusted temperature

$$\text{Prior Day Adjusted Temp}_k = (1 + \text{PDWS})\text{Temp}_k - (\text{PDWS})\text{Temp}_{k-1}. \qquad (3.2)$$

The effect of the prior day adjustment for PDWS $= -0.3$ is demonstrated by Kaefer

in Figure 3.2.

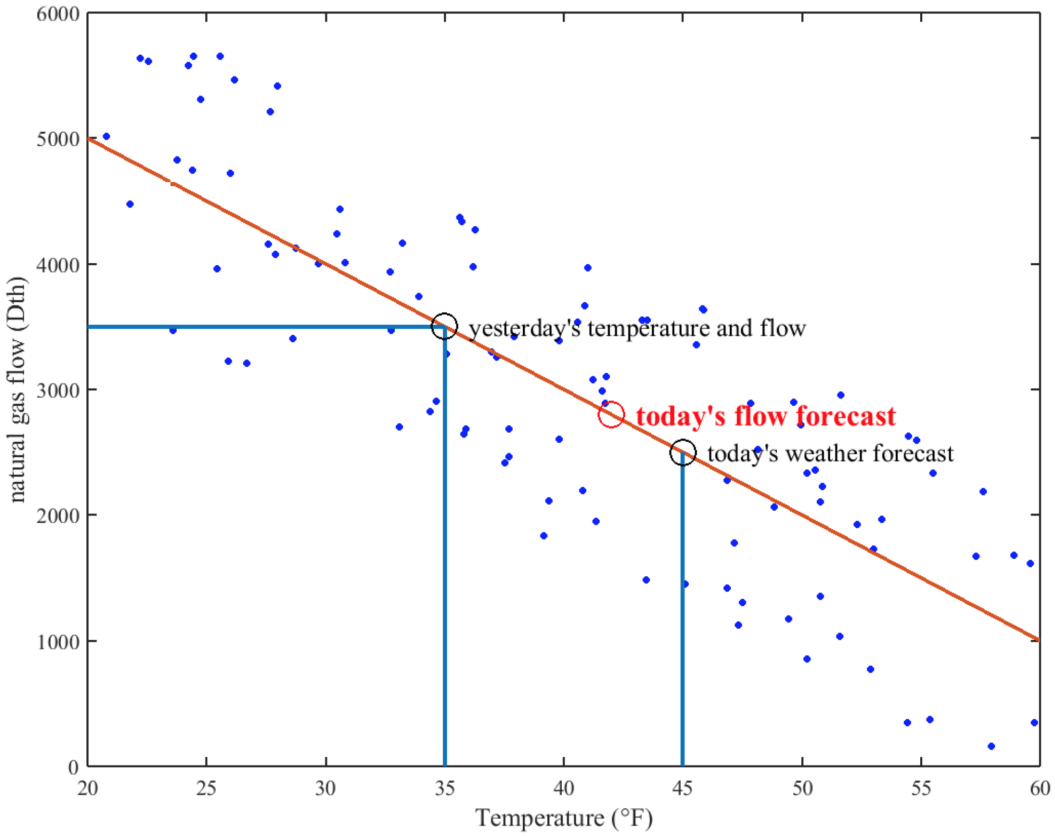


Figure 3.2: Prior Day Adjusted Temperature [24]

Given a PDWS $= -0.3$, the forecast demand is adjusted as if the temperature is 30% of yesterday's temperature and 70% of today's temperature.

Now that we have adjusted temperatures to be better predictors of flow, we can determine the Design Day conditions. The first step to determining the Design Day conditions is to fit distributions to the cold tail of the adjusted temperatures.

3.1.2 Fitting Distributions to Adjusted Temperatures

We fit a Generalized Extreme Value distribution in the Block-Maxima approach using maximum likelihood estimation. We fit a Generalized Pareto

distribution in the Peak Over Threshold approach. We also use the SKDF method described in Section 2.1.1. The SDKF is philosophically a Peak Over Threshold approach because this method fits a kernel density function - a local fit - to temperatures. Descriptions of these distributions can be found in Chapter 2.

To use these distributions we need to set a few hyperparameters manually. To use the Peak Over Threshold approach, we need to set a threshold from which to subtract the peaks. We also need to choose the method of optimizing the parameters of the distribution. For the Surrogate Kernel Density Fit approach, we need to determine the season length and the number of surrogate days to use.

To determine the threshold for the Peaks Over Threshold method, we choose the threshold which is exceeded on average five times per year. Having such a high threshold helps to ensure that exceedences are independent - an ideal characteristic for the Generalized Pareto Distribution [2]. We fit the Generalized Pareto Distribution to these exceedences using method of moments as it has been shown to perform well on data sets with around 100 data points [21]. The method of moments determines the scale σ and shape ξ parameters of a Generalized Pareto distribution. If $\bar{x}$ is the sample mean and s^2 is the sample variance, the scale parameter is

$$\sigma = \bar{x}(\bar{x}^2/s^2 + 1). \tag{3.3}$$

The shape parameter is

$$\xi = \frac{1}{2}(\bar{x}^2/s^2 - 1). \tag{3.4}$$

The Surrogate Kernel Density Fit is dependent on the hyperparameters season length and number of surrogate days to use. D'Silva recommends that we use 91 days. D'Silva claims 91 days is broad enough to include enough days to fit a model too, yet narrow enough to include only the coldest days of the year [11].

Choosing the number of surrogate days to use follows a similar thought process. We want to use as many days as possible, however, a temperature on the 4^{th} of July cannot reasonably be used as a surrogate day for the 1^{st} of January. We have found empirically that 91 days does a reasonable job of creating enough data while using reasonably similar days as surrogates. Using these hyperparameters, the SKDF is fit according to Section 2.1.1.

Next, the Design Day conditions are estimated from each of these three distributions.

3.1.3 Estimating Conditions from Statistical Models

For each of the distributions used, we can determine the One-in-N temperature by taking the inverse of the cumulative density function.

Once a condition is estimated from each distribution, we can create an ensemble condition using combinations of the individual estimates.

To estimate the One-in-N temperature for the Generalized Extreme Value distribution, we evaluate the inverse cumulative density function at $1/N$. To estimate the One-in-N temperature from the SKDF, we take the inverse cumulative density function of $1/(91 \times N)$, since we model 91 days per year. Similarly, for the Generalized Pareto distribution, we take the inverse cumulative density function of $1/(5 \times N)$, since we are modeling 5 days per year.

We ensemble the different predictions by averaging the One-in-N estimates. An average of multiple models adds robustness to the prediction. An averaged prediction *might* not be as good as the best individual estimate, but it will always be better than the worst individual estimate.

Now that we have developed four methods for estimating the Design Day conditions, we forecast the demand corresponding to those conditions.

3.2 Forecasting Design Day Demand

Forecasting Design Day demand introduces two challenges. First, we need to forecast flow where temperature data are particularly sparse; a temperature with a return period of 40 years is not likely to have occurred in a demand dataset with

only 10 years of data. Second, we need to determine the level of confidence we have in our forecasts; what demand are we 99% sure will not be exceeded on the Design Day?

3.2.1 Forecasting Demand During Rare Cold Days

To forecast during extreme cold days, GasDay uses as much data as possible. Because characteristics of an operating area change over time, it is difficult to use all historical data in building a model. GasDay compensates for the changing system by adjusting historical demand to behave similarly to recent demand. GasDay calls this process *detrending* [6].

First we train a linear regression model on the most recent year of data. We then train a model on a previous year of data. Demand on the previous year is adjusted by the difference in the two forecasting models. For example, we train a model on data from 2017

$$\widehat{S}_k = \beta_0^{2017} + \beta_1^{2017}\mathrm{HDD}_k, \tag{3.5}$$

.

where $\widehat{S}_k$ is the forecast demand on day k, and HDD_k is the heating degree day on day k. Similarly, we train a model on data from 2016

$$\widehat{S}_k = \beta_0^{2016} + \beta_1^{2016}\text{HDD}_k. \tag{3.6}$$

.

We then adjust the actual demand on each day 2016 (S_k) by

$$\Delta S_k = \beta_0^{2017} - \beta_0^{2016} + (\beta_1^{2017} - \beta_1^{2016})\text{HDD}_k \tag{3.7}$$

.

resulting in detrended data for 2016. We repeat this for every preceding year in the dataset.

While we can only forecast the Design Day demand from the Design Day condition (an adjusted temperature), we can detrend data using all of the predictors described in [44] and Section 1.3.1. Therefore, we detrend using linear regression models with the following set of inputs

1. Bias
2. HDDW65

3. HDDW55
4. Δ MHDDW - A modified change in HDD from one day to the next
5. CDD65

We can fit a linear regression model to the detrended data. The linear regression model evaluated at the Design Day condition is the expected Design Day demand.

3.2.2 Determining the Level of Confidence in Forecasts

In order for the Design Day demand forecast to be useful, the uncertainty in the forecast must be quantified. Typically, LDCs will ask us to provide 2.5 standard deviations of residuals for our linear fit. For a normal distribution, 2.5 standard deviations above the mean is greater than 99.38% of the density. For this reason we find the linear quantile model that sits above 99.38% of all demand. In order to stabilize our predictions, we perform composite quantile regression on quantiles 99.38, 93.88, and 84.38; the 93.88^{th}, and 84.38^{th} quantiles are used to stabilize the regression - only the parameters for the 99.38^{th} quantile are used.

We also are interested only in the uncertainty during the coldest days. Therefore, we remove all data where the temperature is warmer than 50 degrees. This is motivated by a much tighter bound on demand when there is no heat load.

For this reason, we modify the quantile regression procedure described in Section 2.2.2 to weight more heavily the fit on the coldest days. We rank data points coldest to warmest and train according to the weight shown in Figure 3.3.

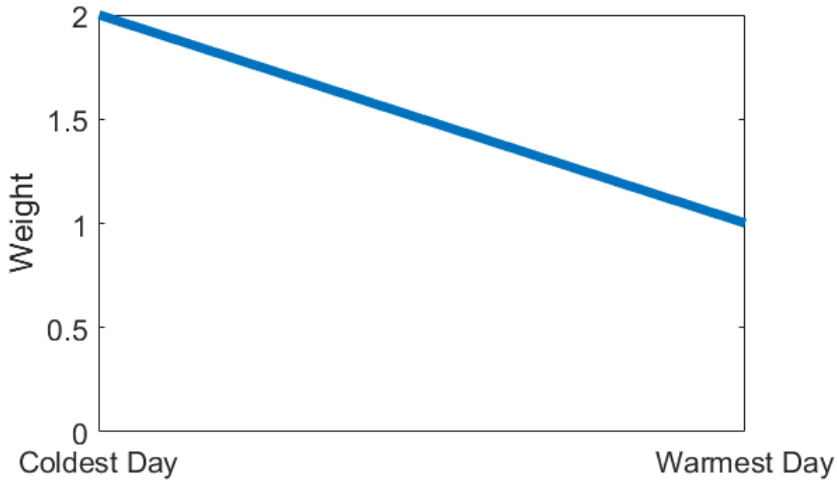


Figure 3.3: Weighting for quantile regression

We weight the cost on quantile regression twice as much on the coldest day in the data set. Then, the cost for all other days is linearly interpolated based on the order of days ranked coldest to warmest.

Now that we have calculated the Design Day conditions and the Design Day demand, we need to evaluate the performance of our forecasts.

3.3 Evaluating Performance of Forecasts

Evaluating the performance of our forecasts is split into two steps. First, we evaluate the performance of our Design Day condition forecast. Second, we evaluate

the performance of our Design Day demand probabilistic forecast. We ignore performance of point forecasts for the Design Day because the focus of this thesis is on probabilistic forecasting of the Design Day.

3.3.1 Evaluating the Design Day Condition Forecast

There are two primary metrics for evaluating the Design Day conditions: volatility and reliability.

First, we will examine the volatility of the forecast (described in Section 2.3.2). We also consider the demonstration of volatility visualized by D'Silva [11] in Figure 3.4.

In practice, LDCs use all previous data to build distributions. Each year, a new year of data is included in building the distribution. Therefore, we will reverse the time axis in Figure 3.4 to emulate the way LDCs will experience changes in forecasts.

We then look at the reliability of our Design Day conditions forecast. We use the actual-vs-expected ratio described in Section 2.3.2. The closer the metric is to 1.0, the more reliable is our forecast.

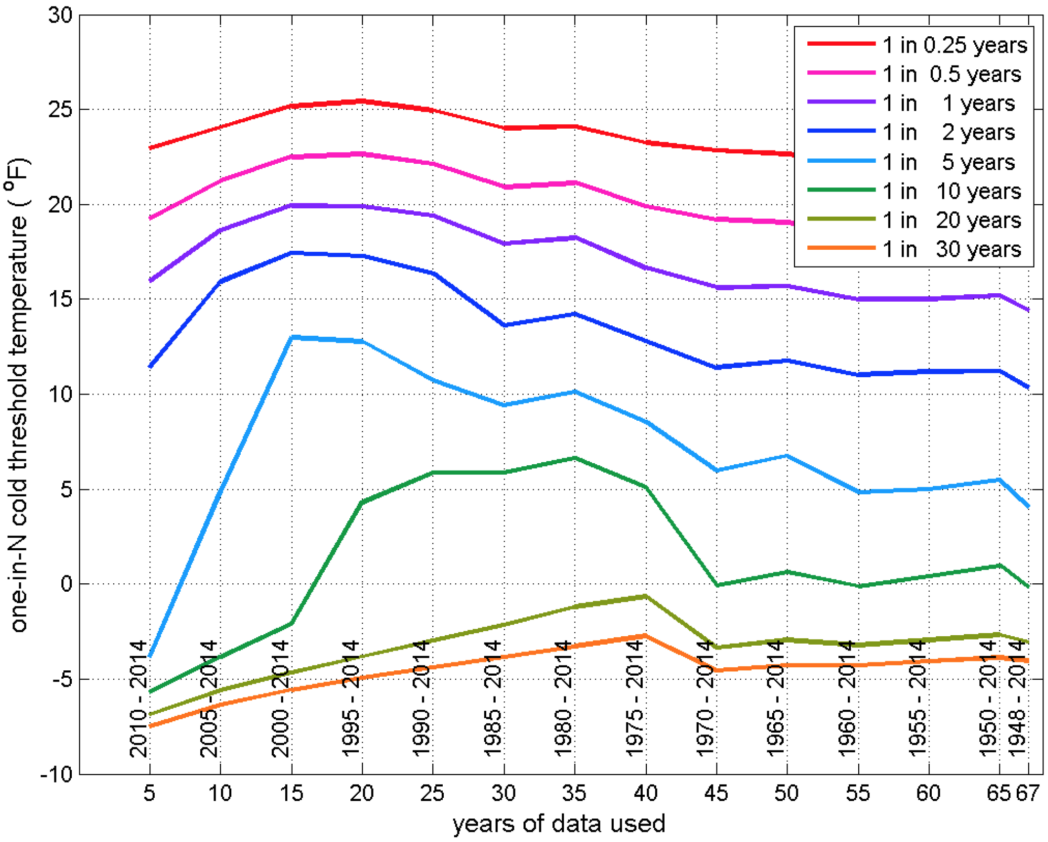


Figure 3.4: Weighting for Quantile Regression

The One-in-N temperature thresholds are plotted over time. Going left-to-right, more data is used for calculating the Design Day conditions. In this chart, data is added by moving the first day of data back. However, in practice, data is added by moving the last day of data forward (as time goes on). We will adjust this plot to better represent what happens in practice.

3.3.2 Evaluating Performance of Design Day Demand Forecast

As mentioned in Section 2.3.3, there are two properties of a probabilistic forecast: *reliability* and *sharpness*. A *reliable* forecast has the expected number of observations between each quantile as the number of observations gets large. *Sharpness* refers to preference towards less uncertainty in a forecast - assuming the

forecast is reliable. Because sharpness is only important if the forecast is reliable, we focus on measuring reliability in this work.

To measure reliability, we perform forecasts on the 99.38^{th} quantile of demand. We perform these forecasts on the 100 most temperature sensitive operating areas for which GasDay has data. The 100 most sensitive operating areas are chosen according to Tenneti [42]. We perform the PIT for the $0.9938 < F(x) < 1$ bin; we count the number of times the actual flow is above the 99.38^{th} quantile for each operating area. The bin should hold 0.62% of the points in each operating area. We then sum across operating areas to get the total number of demands exceeding the 99.38^{th} quantiles and the total expected exceedances. We then calculate the actual-vs-expected ratio.

We also could calculate Saber's metric for reliability [38] - PQRS score,

$$\text{PQRS} = \frac{1}{n}\sum_{i=0}^{n} |\text{Expected percent in bin}_i - \text{Actual percent in bin}_i|. \tag{3.8}$$

n is the number of bins that we are calculating the score for (in this situation, $n = 1$). For $n = 1$, the PQRS is nearly the same metric as the actual-vs-expected ratio; PQRS/Expected percent in bin $= |1 -$ actual-vs-expected ratio$|$. The PQRS

is telling us the same information as the actual-vs-expected ratio. Rather than introducing a new metric to our results, we use actual-vs-expected ratio.

The actual-vs-expected ratio calculates the reliability of quantile forecasts for all operating areas in aggregate. We also consider the scenario that each of the operating area forecasts are biased, but in aggregate the biases cancel out. For example, we expect the demand to be above the 99.38^{th} quantile 10 times across 10 operating areas. If demand is above the 99.38^{th} quantile 10 times in one operating area and zero times in the other operating areas, the actual-vs-expected ratio would be a perfect 1.0. However, these forecasts are clearly flawed. To capture this effect, we calculate the root mean squared error (RMSE) between the actual and expected number of exceedances.

We have described the methods that will be used for estimating the Design Day conditions and Design Day demand. We have discussed methods for evaluating their performance. Next, we perform experiments to determine the performance of our Design Day conditions and Design Day demand estimates in practice.

CHAPTER 4

Evaluation of Our Design Day Analysis

Design Day analysis is split into two steps. First, we predict and evaluate performance of Design Day conditions. Then, we predict and evaluate the performance of Design Day demand. In practice, we would determine the Design Day conditions for an Operating Area and use those conditions to forecast Design Day demand. Long, high quality weather data sets are difficult to procure. For this reason, we perform our analysis of Design Day conditions on many weather stations that are geographically dispersed. We then perform our analysis of Design Day demand.

4.1 Evaluation of Design Day Conditions Forecast

Evaluation of the Design Day conditions is split into two parts. First, we determine the reliability of Design Day condition forecasts made by each of the statistical models described in Table 4.1. Then, we evaluate the volatility of the forecasts to determine how useful they are for practitioners interested in long term planning.

Table 4.1: Summary of methods used in One-in-N experiment

Method	Description
Coldest In Last N Years	Simply use the coldest observed temperature from the last N years as the threshold.
GEV	Make a set containing the coldest day of each year. Fit a GEV distribution to the set.
KDF	Make a set containing the 91 coldest calendar days of each year. Fit a kernel density function to the set.
SKDF	Make a set containing the 91 coldest calendar days of each year. Supplement set with surrogate data method found in Section 2.2.1. Fit a kernel density function to the set.
Generalized Pareto	Fit Generalized Pareto distribution to peaks over cold threshold. Cold threshold is set such that it is exceeded by an average of 2 days each year.
Ensemble	Average of the Generalized Pareto, KDF, and SKDF methods.

4.1.1 Data Source

Hourly temperature and wind data are collected from the National Oceanic and Atmospheric Administration (NOAA) and AccuWeather. The hourly data are averaged into daily data. 38 weather stations, each with 67 years of data were used in this experiment. The stations chosen have high quality data and are geographically diverse; for this reason we refer to these stations as the *continental* dataset. The names of the stations are found in Table 4.2. The locations of these stations are displayed in Figure 4.1.

We also use anonymous temperature and wind data from an LDC in the

Table 4.2: Stations in *continental* dataset

Station Location	Callsign	Station Location	Callsign
Camp Springs, MD	KADW	Dallas-Fort Worth, TX	KDFW
Amarillo, TX	KAMA	Aspen, CO	KASE
Boston, MA	KBOS	Seattle, WA	KSEA
Brownsville, TX	KBRO	New York, NY	KNYC
Corpus Christi, TX	KCRP	Minneapolis, MN	KMSP
Dayton, OH	KFFO	Miami, FL	KMIA
Fort Smith, AR	KFSM	Bakersfield, CA	KBFL
New Orleans, LA	KMSY	Calgary, AB	CYYC
Pittsburgh, PA	KPIT	Winnipeg, MB	CYWG
Pueblo, CO	KPUB	Vancouver, BC	CYVR
Raleigh/Durham, NC	KRDU	Regina, SK	CYQR
Riverside, CA	KRIV	Ottawa, ON	CYOW
San Antonio, TX	KSAT	Memphis, TN	KMEM
Salt Lake City, UT	KSLC	Jackson, MS	KJAN
Tulsa, OK	KTUL	Nashville, TN	KBNA
Valparaiso, FL	KVPS	Kansas City, MO	KMCI
Wrightstown, NJ	KWRI	Hays, KS	KHYS
King Salmon, AK	PAKN	Evansville, IN	KEVV
Honolulu, HI	PHNL	Louisville, KY	KSDF

southeast United States. This LDC uses 45 stations to determine their Design Day conditions. There is not as much geographic diversity as we see in Table 4.2. However, it provides a case-study and shows how these methods perform for an actual LDC. These stations are referred to as the *case-study* stations.

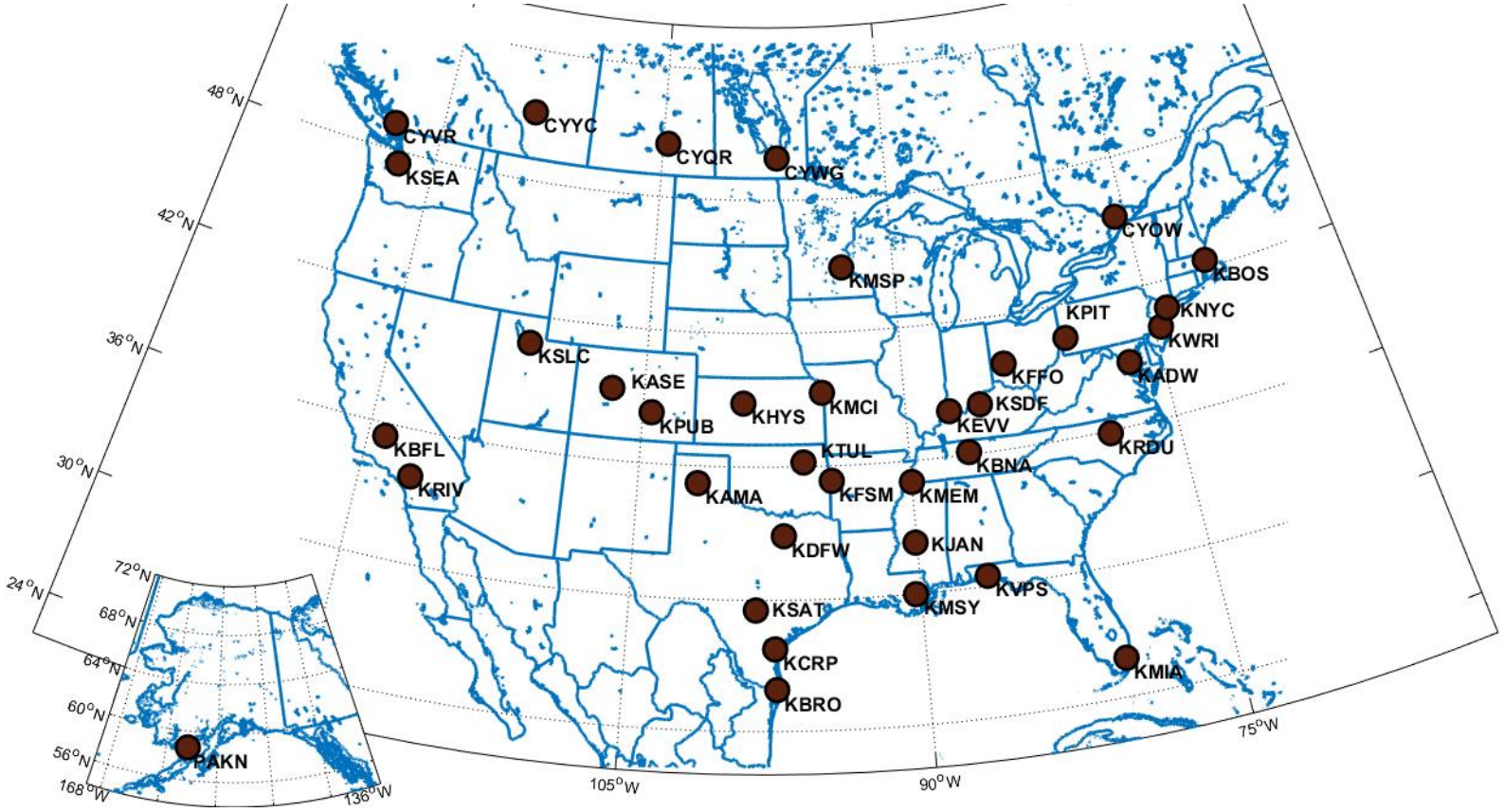


Figure 4.1: Stations in the *continental* dataset

The map shows the geographic diversity of stations used in the United States and Canada. The four letter code is known as the Callsign - a unique identifier of the station. Omitted from this map is PHNL, or Honolulu HI.

4.1.2 Experiment

Ideally, we could fit distributions to thousands of years of stationary temperature data. Unfortunately, we are limited to 67 years of data, and climate is not stationary. We approach this challenge by testing our One-in-N forecast in two ways.

First, we perform an in-sample test. For each station, we estimate the One-In-N condition using all 67 years of data. We then count the number of temperatures colder than the threshold in all 67 years of data. The actual-vs-expected ratio (described in Section 2.3.2) is calculated. An in-sample test is a recommended method in climate statistics when the amount of data required to make the prediction is roughly the amount of data that exists [45].

We also perform an out-of-sample test. In this case, we determine the One-in-N threshold from all but one year of data. We then count the number of temperatures colder than the One-in-N threshold in the held-out year. We repeat this by holding out each year. This results in 2546 tests (67 held out years $\times$ 38 weather stations). We aggregate all of these tests to calculate the actual-vs-expected ratio.

These tests are run on both the stations from Table 4.2 and the anonymous stations from the *case-study*. The *case-study* stations are only tested using raw

temperature to keep the results section succinct. The stations from Table 4.2 are tested using raw, wind-adjusted, and prior-day adjusted temperatures - starting with raw temperatures.

Using Raw Temperature: *Continental* Dataset

The experiment is run on the raw temperatures from each of the weather stations. The in-sample actual-vs-expected ratios are provided in Table 4.3.

We want these ratios to be near 1.0. A ratio greater than 1.0 means that the estimator was biased too warm; more temperatures than expected are colder than the threshold. A ratio less than 1.0 means that the estimators are biased too cold.

The table contains ratios for the One-in-10, 20, 30, 40 conditions. The variance of a distribution sampled in the extreme tails is relatively high, so we put less emphasis on the actual-vs-expected ratio for larger N.

Table 4.3: *Continental* in-sample actual-vs-expected ratio raw temperature

N	**Coldest In Last N Years**	**GEV**	**KDF**	**SKDF**	**Generalized Pareto**	**Ensemble**
10	3.418	1.75	0.799	0.931	1.141	0.908
20	5.75	1.622	0.776	0.854	0.954	0.838
30	1.35	1.501	0.687	0.885	0.954	0.896
40	1.148	1.381	0.667	0.916	0.962	0.776

For the in-sample test, the SKDF and the Generalized Pareto methods performed the best and exhibited little bias. The GEV method is biased too warm. The GEV fits a distribution to the coldest day of each year. Because temperature is auto-correlated, we might expect the coldest two temperatures in history to occur in the same year. Therefore, the GEV is not fit to some of the coldest days, explaining its bias towards being too warm. The Coldest In Last N Years is an unreliable estimator. It performs reasonable well for large N, but this is more of a problem with the test than a reflection of the method's performance. For example, if we chose N = the number of years in the dataset, the Coldest In Last N Years method would perform perfectly on the in-sample test.

The in-sample fit for each of the distributions is shown in Figure 4.2. Four stations of the 38 were selected at random for the figure. Also labeled is the threshold for the One-in-30 condition. All four of the distribution-based methods place the One-in-30 condition is a similar location, lending confidence that each of these methods are reasonable estimators of the Design Day condition. The GEV fit does not appear to accurately represent the in-sample data, which also contributes to the poor actual-vs-expected ratio.

The out-of-sample test shows similar results. The SKDF and the KDF perform the best. The Generalized Pareto method sets the threshold slightly too

Table 4.4: *Continental* out-of-sample actual-vs-expected ratio raw temperature

N	**Coldest In Last N Years**	**GEV**	**KDF**	**SKDF**	**Generalized Pareto**	**Ensemble**
10	3.471	1.812	0.994	1.057	1.25	1.061
20	5.904	1.8	0.959	1.006	1.195	0.991
30	1.651	1.84	0.967	1.026	1.262	1.05
40	1.211	1.714	0.975	1.069	1.289	1.132

warm; it is exceeded more often than expected. The GEV and the Coldest In Last N Years perform similarly to the in-sample test.

Table 4.5: *Continental* volatility of raw temperature threshold

N	**Coldest In Last N Years**	**GEV**	**KDF**	**SKDF**	**Generalized Pareto**	**Ensemble**
10	3.6023	0.0507	0.053	0.0557	0.0623	0.052
20	2.1245	0.089	0.0753	0.0729	0.0979	0.071
30	0.4952	0.126	0.0964	0.0872	0.1277	0.0856
40	0.2246	0.0635	0.0452	0.0422	0.0899	0.0432

Table 4.5 contains the volatility of each method. A volatility near 0.0 represents a method that is useful to an LDC; the Design Day condition changes slowly over time. All of the distribution-based methods perform well relative to the Coldest In Last N Years method. The volatility is visualized in Figure 4.3. Most of the methods vary slowly over time. The Coldest In Last N Years method is constant most years, but occasionally makes a huge jump. When an extreme cold event occurs, the Coldest In Last N Years threshold steeply drops. When an extreme cold event falls out of the last N years window, the Coldest In Last N Years threshold steeply rises. This flaw in the Coldest In Last N Years method has been a complaint of many LDCs that work with GasDay.

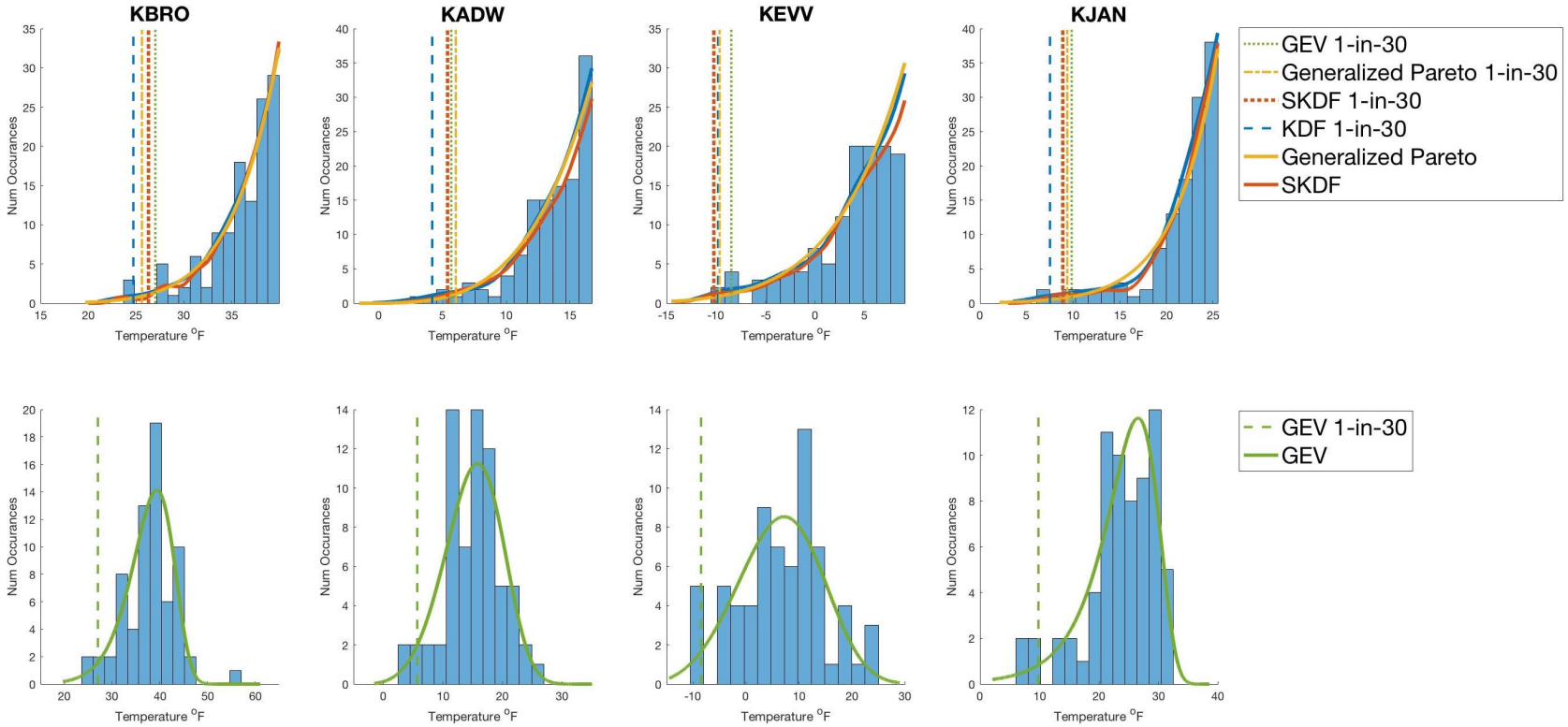


Figure 4.2: *Continental* raw temperature distribution fits

These charts show the fits of distributions to the coldest temperatures. The top row shows the fits of the Generalized Pareto, SKDF, and KDF. The bottom row shows the fit of the GEV to the coldest day of each year. The figures for all stations in the *continental* dataset can be found in Appendix A.

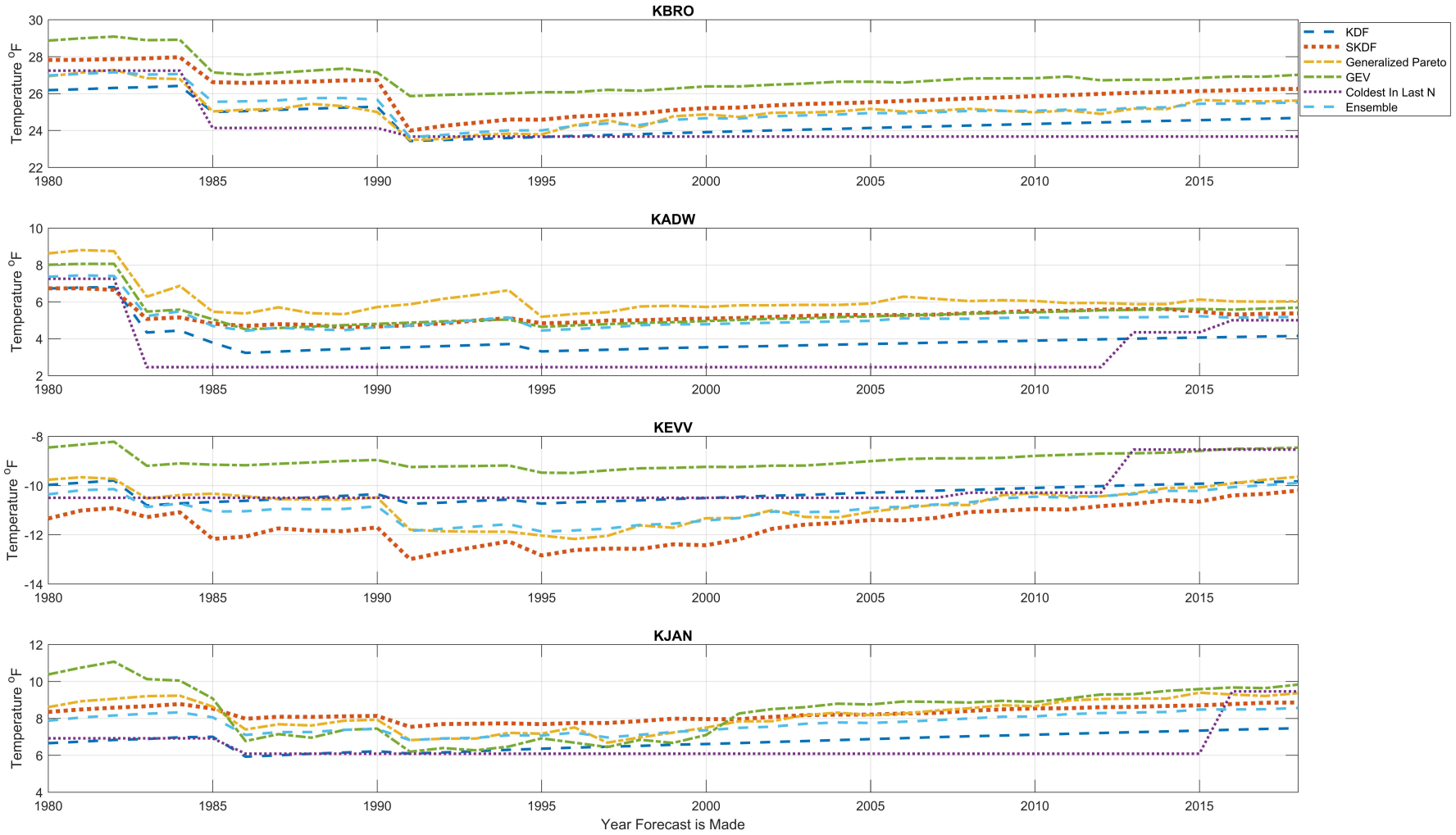


Figure 4.3: *Continental* changes in One-in-30 condition over time

Using Raw Temperature: *Case-study* Dataset

While we should have more confidence in the *continental* results, it is important for a method to perform well on the *case-study* dataset; these results are delivered to LDCs as justification for using one of these methods.

Table 4.6: *Case-study* in-sample actual-vs-expected ratio raw temperature

N	**Coldest In Last N Years**	**GEV**	**KDF**	**SKDF**	**Generalized Pareto**	**Ensemble**
10	2.435	1.786	0.926	1.099	1.063	1.03
20	4.838	1.486	0.841	0.971	0.919	0.88
30	1.252	1.291	0.743	0.968	0.88	0.89
40	1.004	1.2	0.795	1.017	0.848	0.821

The in-sample results provide justification for using the SKDF, Generalized Pareto, and Ensemble methods. All have actual-vs-expected ratios relatively near 1.0 (see Table 4.6). The out-of-sample test shows each method to be biased slightly too warm; the thresholds are exceeded too often (see Table 4.7). However, since the SKDF, Generalized Pareto, and Ensemble methods are all still relatively close to 1.0, these methods are still justifiable.

Calculating volatility is perhaps the most important test for the *case-study* dataset; it directly assesses the usability of each method for an actual LDC. In

Table 4.7: *Case-study* out-of-sample actual-vs-expected ratio raw temperature

N	**Coldest In Last N Years**	**GEV**	**KDF**	**SKDF**	**Generalized Pareto**	**Ensemble**
10	3.101	1.926	1.059	1.225	1.205	1.122
20	6.062	1.733	1.042	1.235	1.248	1.129
30	1.624	1.614	1.026	1.225	1.195	1.155
40	1.195	1.58	1.076	1.275	1.288	1.195

Table 4.8 we find very similar results to the *continental* threshold volatility. All methods have low volatility except for the Coldest In Last N Years method.

Table 4.8: *Case-study* volatility of raw temperature threshold

N	**Coldest In Last N Years**	**GEV**	**KDF**	**SKDF**	**Generalized Pareto**	**Ensemble**
10	4.746	0.08	0.0946	0.089	0.1054	0.0891
20	3.6321	0.1585	0.2027	0.1336	0.1668	0.1466
30	1.1162	0.2447	0.2444	0.2127	0.2191	0.1959
40	0.3134	0.152	0.0644	0.0607	0.1207	0.0599

Using the *case-study* dataset is a useful demonstration of how each method works for an actual LDC. However, we shift our focus back onto the *continental* dataset, as it demonstrates performance across a larger geographic region. Analyzing both datasets is too cumbersome as we analyze performance on wind and prior day adjusted temperatures.

Using Wind Adjusted Temperatures: *Continental* Dataset

The wind adjustment to temperatures is made to create a better predictor for flow. We perform the same test for determining the Design Day condition of a wind adjusted temperature as we did for the raw temperature. The in-sample results again show the Generalized Pareto and SKDF methods perform the best. The KDF is still biased too cold, and the GEV is still biased too warm.

Table 4.9: *Continental* in-sample actual-vs-expected ratio wind adjusted

N	**Coldest In Last N Years**	**GEV**	**KDF**	**SKDF**	**Generalized Pareto**	**Ensemble**
10	3.349	1.638	0.835	0.987	1.106	0.999
20	5.839	1.507	0.852	1.032	1.065	0.942
30	1.376	1.339	0.835	1.032	1.056	0.958
40	1.13	1.261	0.803	1.015	1.097	0.934

The out-of-sample test yields very different results after adjusting for wind. Here, the KDF performs best; the SKDF, Generalized Pareto, and Ensemble are slightly biased warm. While all methods perform reasonably, the change in bias shows that wind adjusting the temperatures seems to increase the width of the distribution tails. This is best examplified when comparing KADW in Figure 4.4 to KADW in Figure 4.2. Clearly, the tail is wider in the wind adjusted distribution, and the distribution fits fail to capture the thicker tail.

Table 4.10: *Continental* out-of-sample actual-vs-expected ratio wind adjusted

N	**Coldest In Last N Years**	**GEV**	**KDF**	**SKDF**	**Generalized Pareto**	**Ensemble**
10	3.299	1.71	1.021	1.095	1.249	1.137
20	5.701	1.693	1.004	1.162	1.278	1.129
30	1.593	1.618	1.046	1.232	1.332	1.133
40	1.145	1.627	1.162	1.261	1.378	1.212

Table 4.11 - representing the volatility in the wind predictions - does not provide any new information. We see that all methods other than Coldest In Last N Years perform reasonably.

Table 4.11: *Continental* volatility of wind adjusted temperature threshold

N	**Coldest In Last N Years**	**GEV**	**KDF**	**SKDF**	**Generalized Pareto**	**Ensemble**
10	5.2986	0.0751	0.0915	0.073	0.0888	0.0776
20	3.7863	0.1381	0.1208	0.1043	0.1454	0.1092
30	1.1386	0.1742	0.1359	0.1272	0.1799	0.1239
40	0.7718	0.0841	0.0355	0.0692	0.1059	0.0499

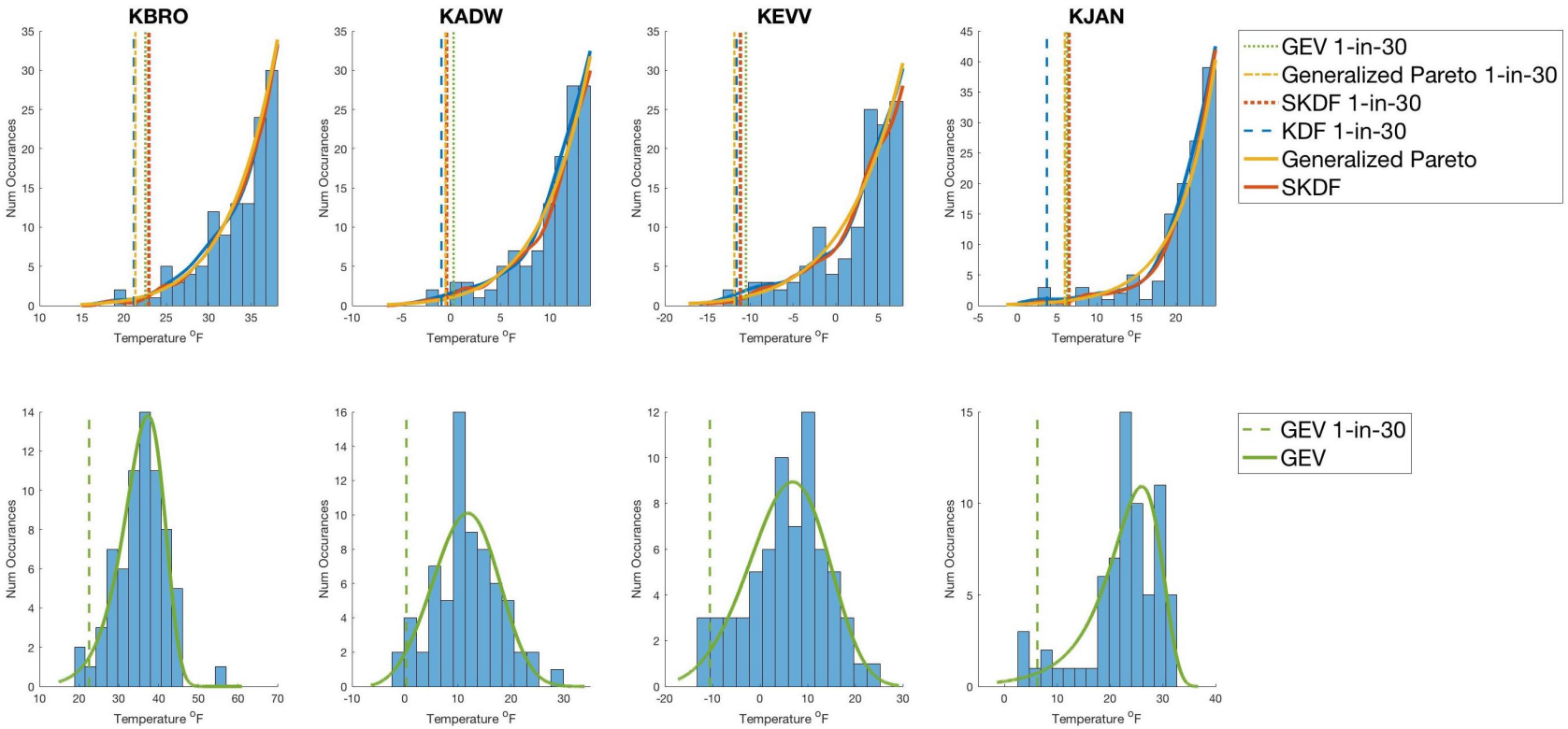


Figure 4.4: Wind adjusted temperature distribution Fits

Using Prior Day Adjusted Temperatures: *Continental* Dataset

We adjust temperatures based on the prior day temperature to improve the correlation with demand. As described in Figure 1.3, demand data is required to calculate the prior day effect. However, this experiment is conducted independent of demand. Therefore, we calculate the prior day effect from one operating area and apply it to all stations in our test.

Table 4.12: *Continental* in-sample actual-vs-expected ratio prior day adjusted

N	**Coldest In Last N Years**	**GEV**	**KDF**	**SKDF**	**Generalized Pareto**	**Ensemble**
10	4.504	1.936	0.842	0.927	1.071	0.927
20	6.665	1.87	0.737	0.861	1.009	0.815
30	1.49	1.757	0.605	0.78	0.908	0.71
40	1.117	1.474	0.636	0.714	0.869	0.714

According to the in-sample test (found in Table 4.12), the Generalized Pareto and SKDF methods perform the best. The KDF, SKDF, and Ensemble are all biased too cold, especially for large N. This is reflected by the significantly thinner tail (particularly for KADW) in Figure 4.5.

The KDF, SKDF, and Ensemble perform best for the out-of-sample test

Table 4.13: *Continental* out-of-sample actual-vs-expected ratio prior day adjusted

N	**Coldest In Last N Years**	**GEV**	**KDF**	**SKDF**	**Generalized Pareto**	**Ensemble**
10	4.465	1.993	0.979	1.034	1.215	1.065
20	6.682	2.02	1.006	1.006	1.289	1.108
30	1.675	2.04	0.991	1.108	1.486	1.156
40	1.164	1.997	1.006	1.069	1.447	1.038

(shown in Table 4.13). The Generalized Pareto and - to a greater extent - the GEV methods are biased too warm. The Coldest In Last N Years method again proves to be an unreliable method.

Table 4.14: *Continental* volatility of prior day adjusted temperature threshold

N	**Coldest In Last N Years**	**GEV**	**KDF**	**SKDF**	**Generalized Pareto**	**Ensemble**
10	3.1095	0.0427	0.0522	0.0693	0.0654	0.0566
20	2.0736	0.0742	0.0642	0.0833	0.0984	0.0694
30	0.3549	0.1095	0.0747	0.0891	0.1253	0.0782
40	0.243	0.0741	0.0403	0.0383	0.0882	0.0414

Table 4.14 shows the volatility of the different methods applied to prior day adjusted weather. Again, we see that all methods except the Coldest In Last N Years perform reasonably well - implying these methods are usable in practice.

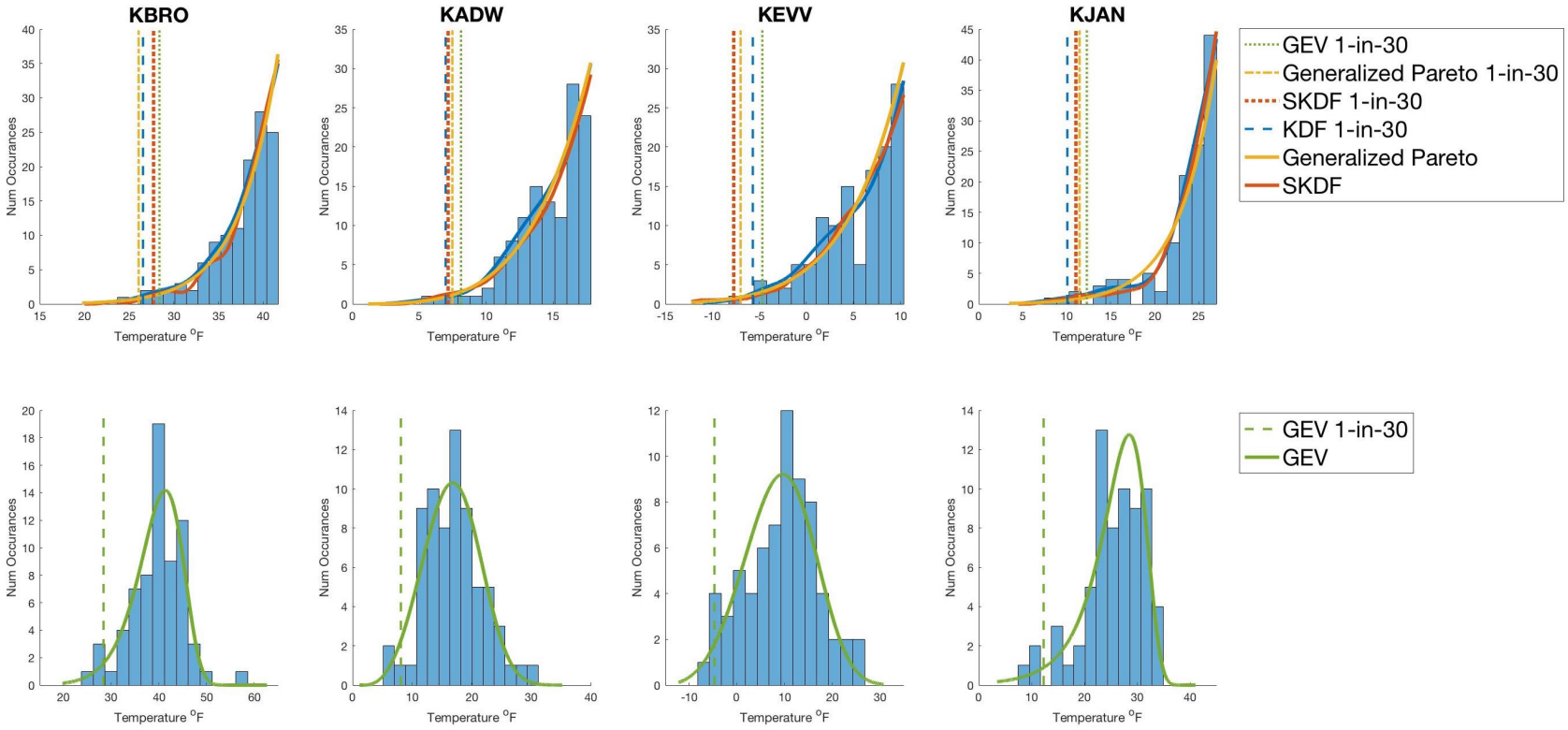


Figure 4.5: Prior day adjusted temperature distribution fits

4.1.3 Discussion

Considering the results in aggregate, the KDF, SKDF, Generalized Pareto, and the Ensemble methods all prove to be reasonably reliable methods for estimating the Design Day conditions. These four methods estimate very similar temperatures for Design Day conditions, and their probability distribution functions are very similar (see Figures 4.2, 4.4, 4.5). It is encouraging to see the methods that performed well empirically (the SKDF and KDF) estimate similar Design Day conditions as the method that has much theoretical backing (the Generalized Pareto).

The GEV method does not perform particularly well on most of the tests - however, most of the temperature estimates are within a couple of degrees of the other methods. The GEV method fundamentally is estimating a different temperature than what we are testing for. The Block-Maxima approach discards all but one datapoint in every year. A set containing the coldest temperature in each of the 30 years does not necessarily contain - and due to autocorrelation is *unlikely* to contain - the coldest 30 temperatures in the last 30 years. Since the coldest days are not necessarily being modeled, it is no surprise the the GEV is biased warm.

Using the Block-Maxima approach may be useful depending on the question to be answered. If the question is “what is the coldest temperature in a year that

we expect to be exceeded once every N years," then the Block-Maxima approach is correct. However - in our test - we asked the question "what is the temperature we expect to be exceeded - on average - once every N years?" The difference is subtle, but explains the poor performance by the GEV in our experiment.

For a fair test, we recalculate the actual-vs-expected ratio using the number of years in which the GEV threshold is exceeded (rather than the total days that exceed the GEV threshold). The recalculated actual-vs-expected ratios using non-adjusted temperature are found in Tables 4.15 and 4.16. Though still biased warm, the actual-vs-expected ratios are much closer to 1.0; the GEV method is reasonable so long as we want to know temperature that will be exceeded in one year out of N.

Table 4.15: *Continental* in-sample Block-Maxima actual-vs-expected ratio

N	**GEV**
10	1.04
20	1.195
30	1.269
40	1.21

Table 4.16: *Continental* out-of-sample Block-Maxima actual-vs-expected ratio

N	**GEV**
10	1.053
20	1.226
30	1.45
40	1.494

The Coldest In Last N Years method performs the worst. Not only does it perform poorly on the actual-vs-expected ratio metric, but the volatility of the method makes it difficult for practitioners to use. For these reasons, using the Coldest In Last N Years method to determine Design Day conditions is not recommended.

4.2 Evaluation of Design Day Demand Forecast

We now shift our focus to forecasting demand given Design Day conditions. The Design Day demand forecast is used by practitioners to make sure they are able to meet demand during a Design Day. We need to capture the uncertainty in the Design Day demand forecast. In particular, we need to determine the flow that will not be exceeded 99.38% of the time on the Design Day. This is particularly difficult to evaluate performance because there is not much demand data on days similar to the Design Day. Further, there is not much demand data that exceeds the 99.38^{th} quantile. In this way, we are looking at the tails of two distributions; we are looking

at the extreme cold temperatures; we are looking at the extreme high demand conditional on an extreme cold temperature. Data is extremely sparse in these two extremes. To compensate, we use 100 anonymized operating areas and run our test in-sample.

4.2.1 Data Source

The 100 most temperature sensitive operating areas in the GasDay data repository are used in this experiment. Temperature sensitivity is calculated according to Tenneti in his thesis [42].

4.2.2 Experiment

The quantile regression model described in Section 3.2.2 is compared to a baseline model. For the baseline, we use a linear regression model with inputs HDDW65, HDDW55, CDD65, and a bias term. We calculate the standard deviation of error on the 20% coldest days. Adding 2.5 standard deviations to the linear regression model provides the 99.38^{th} quantile, assuming errors are normally distributed.

We calculate the actual-vs-expected ratio for each operating area; we limit the testing set to the 10% coldest days, then we count the number of points above the 99.38^{th} quantile. We know that this includes days much warmer than the Design

Day, but this much data is required for analysis. We aggregate the number of expected exceedances and actual exceedances across operating areas and calculate the actual-vs-expected ratio. We also calculate the RMSE for actual and expected exceedances across operating areas, as described in Section 3.3.2.

According to Table 4.17, both methods exhibit little bias when aggregated across the 100 operating areas; the actual-vs-expected ratio is near 1.0 for each method. The composite quantile regression method has a much lower RMSE than the baseline model, indicating a higher resolution; for each station, the number of actual exceedances is closer to expected.

Table 4.17: Uncertainty results for quantile regression and baseline model

Method	**RMSE**	**actual-vs-expected ratio**
Linear With Normal Distribution	3.0338	1.0985
Quantile Fit	1.7249	1.022

Four example fits are shown in Figure 4.6. Though we are only interested in the 99.38^{th} quantile, we also include the 0.62^{th} quantile. This is done to visualize the change in uncertainly with respect to temperature. Figure 4.6.d demonstrates uncertainty decreasing as temperature decreases; the difference between the high and low quantiles decreases as temperature decreases. To contrast this, in Figures

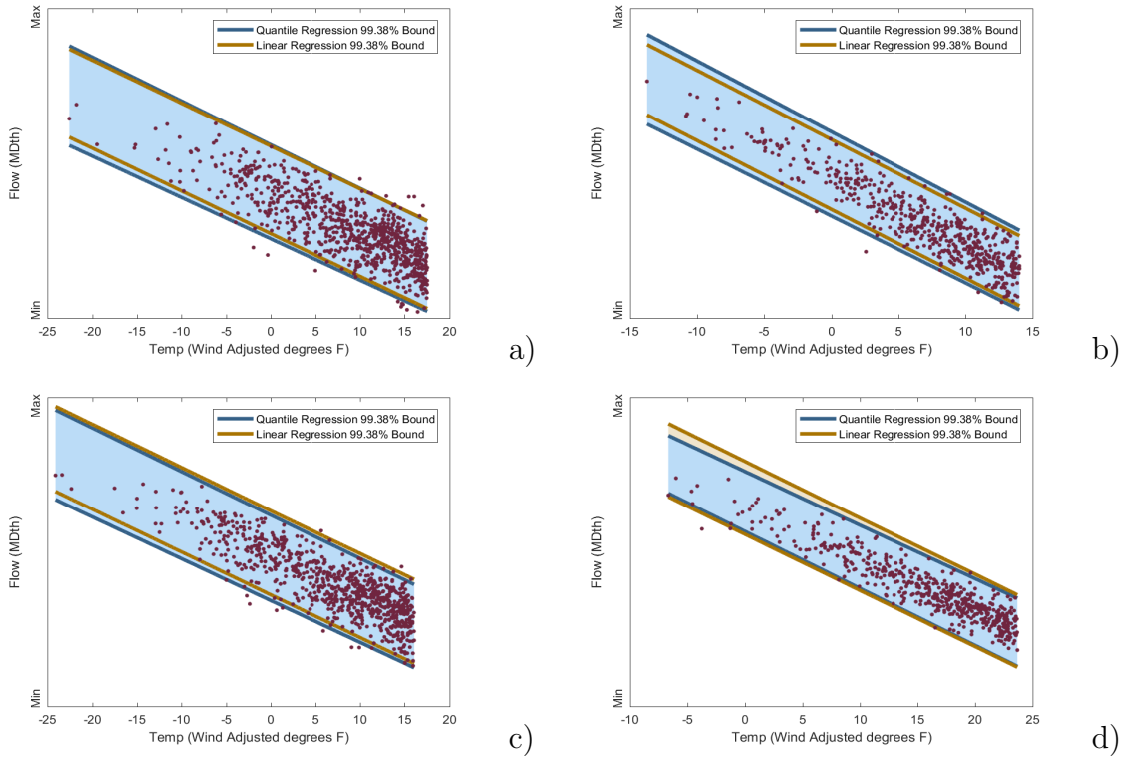


Figure 4.6: Quantiles for four of the most temperature-sensitive operating areas

The figures above show the 99.38% and 0.62% quantiles for four anonymized operating areas. The figures are zoomed into the 10% coldest days. The golden lines represent the baseline model quantiles. The blue lines represent the composite quantile regression quantiles. The blue area represents the region in which we expect 98.76% of points to lie (99.38% - 0.62%).

4.6.a, 4.6.b, and 4.6.c, uncertainty increases as temperature decreases. The baseline model is constrained to have constant uncertainty relative to temperature. Relaxing this constraint allows the composite quantile regression method to better fit the high and low quantiles of the data.

4.2.3 Discussion

While both the quantile fit and the linear regression with normal distribution prove to be reliable - the actual-vs-expected ratios are near 1.0 - the composite quantile regression method proves to be superior in resolution. Both methods can be used with empirical justification. However, the composite quantile regression model is recommended.

There are two major shortcomings with this test, both stem from the lack of data. First, the 10% coldest days in the datasets are much warmer than the Design Day conditions. Our metrics do not truly represent performance on the Design Day. Second, there is no out-of-sample test. Design Day analysis is used for future planning. Ideally, we would have enough data to test several held out years.

There are other challenges that come with forecasting uncertainty in the next year's Design Day. We need to determine how much uncertainty is caused by potential changes in the region being forecasted. New houses may be built, or old houses may be made more efficient. These demographic changes will certainly affect the relationship between temperature and flow. Saber dealt with this by basing his probabilistic forecast on out-of-sample errors [38].

However, the use of nearest neighbors in Saber's method limits the ability to extrapolate - a necessary feature for predicting uncertainty during extreme cold days. This option, among other ideas, are discussed next in the Future Considerations section.

CHAPTER 5

Benefits of Design Day Analysis and Future Considerations

In this chapter we discuss our contributions, improvements that can be made to our Design Day analysis, and extensions to this research outside of the scope of Design Day analysis.

5.1 Contributions of Design Day Analysis

This thesis contributes to Design Day analysis in three ways. First, we improve on current methods for estimating Design Day conditions and provide an out-of-sample analysis. Second, we improve methods for determining Design Day demand. Finally, we improve on methods evaluating performance of Design Day analysis.

Estimating Design Day conditions has been the target of research [10, 11, 25, 31]. This thesis is the first time the empirically good SKDF is compared to - and ensembled with - the Generalized Pareto distribution. Several distributions - both empirically and theoretically based - estimate similar Design Day conditions; we have increased confidence that each of these methods is a reasonable estimator of Design Day conditions.

Estimating the Design Day demand has also been a target of research [3, 4]. This is the first research focused on forecasting uncertainty in our Design Day demand prediction. We also apply a weighting to the training examples during optimization - a novel adjustment to composite quantile regression. We show improvement over the baseline model on an in-sample test.

We introduce the test for volatility in Design Day analysis. It addresses the concern about Design Day conditions that GasDay hears most often from LDCs; a huge change in the Design Day conditions from one year to the next makes it difficult to plan for the future. By running this test on several methods, we show which methods are useful to LDCs. In particular, the KDF, SKDF, Generalized Pareto, and Ensemble methods all perform reasonably well and should be considered as options for estimating Design Day conditions.

5.2 Future Improvements to Design Day Analysis

In this section, we discuss improvements that can be made to estimating Design Day conditions, Design Day demand, and evaluating the performance of our estimates.

In our work estimating the Design Day conditions, we assumed climate to be constant. Huang et al. show that the extreme cold days are expected to occur less

frequently due to climate change. They also show in increase in variance of extreme cold days [22]. To address this, we can detrend the data to remove any trend of rising temperatures. By doing this we can treat all temperature data as if it came from the most recent year. This is particularly challenging because of the large variability of temperatures each year - particularly when we look at a single weather station. Because temperatures vary so much from one year to the next, it is difficult to identify long term temperature trends due to climate change. Determining the long term temperature trend is the main challenge in this future work.

Our probabilistic forecast of Design Day demand does not account for the potential changes in an operating area from one year to the next. This can yield forecasts with too narrow of confidence bounds. To address this, we can implement a section of Saber's method. Saber uses the error of a linear regression model as a basis for his probabilistic forecast. The sets of error he uses are determined from a year of held out demand data. Systematic changes to the operating area being modeled contribute to the errors. Therefore, by modeling the errors, we are modeling the uncertainty caused by potential changes in the operating area from year to year.

Saber's method was not used in this thesis because it creates a probabilistic forecast similarly to K-nearest-neighbors; it can only interpolate, it can not extrapolate. By combining the methods used by Saber and composite quantile

regression, we can solve the problems of both methods; the combination would be able to extrapolate with respect to temperature and be adjusted based on the potential for changes to the operating area being modeled.

Finally, an out-of-sample test would be useful for the Design Day demand. After incorporating the uncertainty that comes with a changing operating area, an out-of-sample test would be useful. The primary challenge is dealing with the sparsity of data. To compensate for this, we can use Kaefer's methods for surrogate data [24]. Surrogate data can be used to increase the size of the out-of-sample test set. With a larger out-of-sample test set, there will be enough data in the tails of the temperature and flow distributions to perform a coherent experiment.

5.3 Future Work

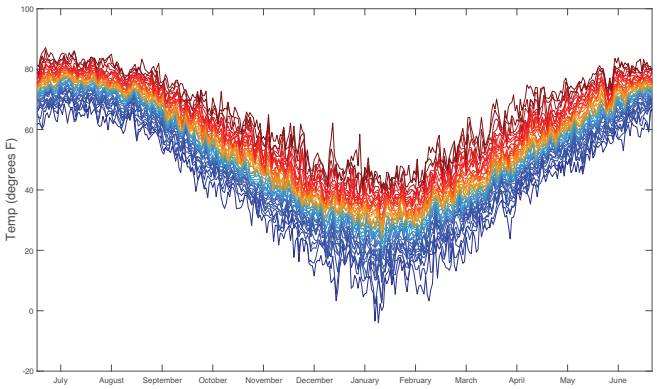
There are two primary suggestions for future work outside of Design Day analysis.

1. Combine composite quantile regression and the surrogate temperature method used in the SKDF method for long term probabilistic forecasts of natural gas demand.
2. Combine Saber's method, composite quantile regression, and probabilistic weather forecasting.

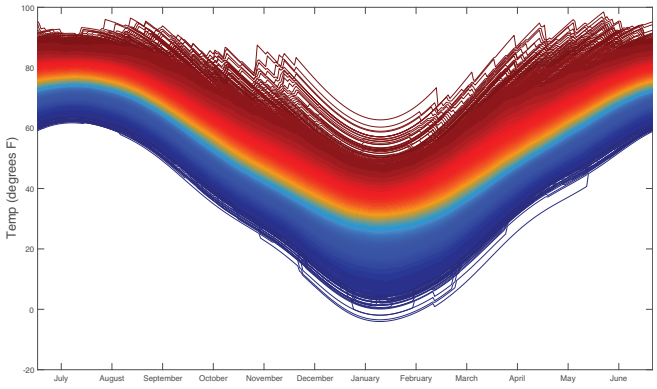
5.3.1 Monthly Probabilistic Forecasts

LDCs need to purchase gas months in advance to mitigate risk. Without reliable weather forecasts, there is little they can do to determine how much gas will be demanded months in advance. LDCs can look at the amount of demand used in previous years, but many operating areas have only been collecting data for the past few years. By modeling the probability distributions of temperatures across the year, and the probabilistic relationship between flow and temperature, we can estimate the probability distributions of demand on each day of the year.

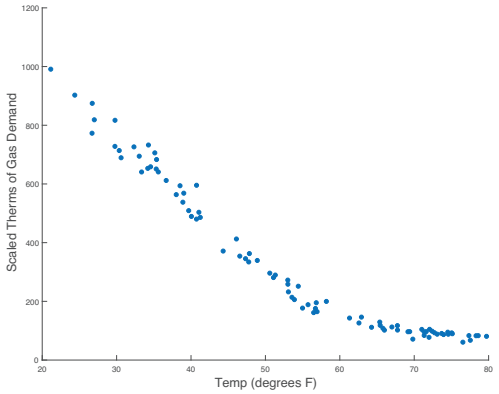
Many of the LDCs that work with Marquette University GasDay^{TM} are more interested in the total amount of demand in a month - particularly when doing the long term forecasting described here. We therefore average the surrogate temperatures into monthly temperatures and build multiple quantile regression models (one for each quantile 1-99) on average monthly demand and temperatures. By using the distribution of monthly temperatures as the input to the quantile regression model, we can estimate a distribution of demand.



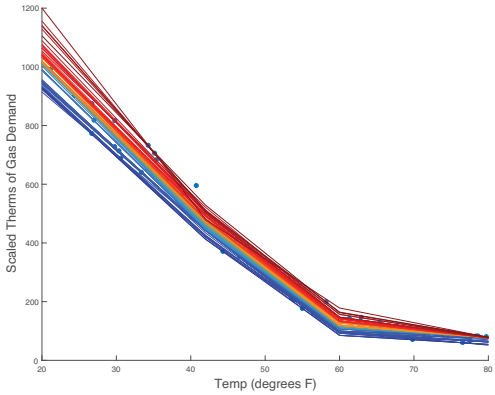
A) The subfigure above shows the temperatures on each day of the year. The temperatures on each day of the year are sorted, then plotted in the same series as the temperatures of the same rank that occured on other days of the year.



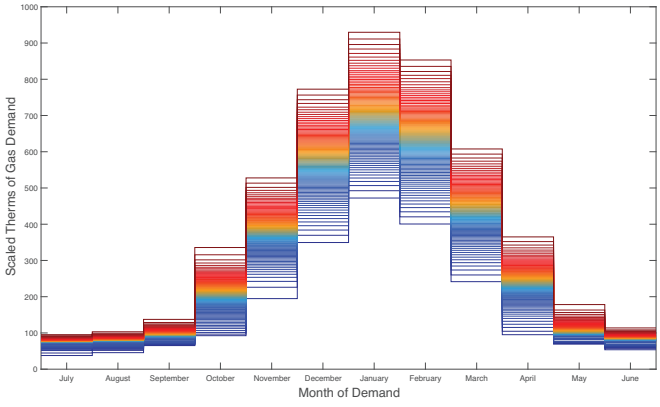
B) The surrogate data process used in the SKDF is applied to the temperatures in A. The subfigure above is analogous to A, but includes the surrogate data



C) Monthly demand (y-axis) is plotted against monthly temperatures (x-axis)



D) Quantile regression models are fit to the data in C. Each line represents a quantile fit. The red lines represent high quantiles, while the blue lines represent the low quantiles.



E) The surrogate temperatures from B are aggregated into monthly temperatures. The monthly temperatures are input into the quantile regression models in D. The resulting probability distribution of demand for each month is plotted above.

Figure 5.1: Monthly Flow Forecast

5.3.2 Daily Probabilistic Forecasts

Daily probabilistic gas demand forecasts is an area of research interest at Marquette University GasDay™. Marquette University GasDay™ provides daily demand forecasts to utilities across the United States. To improve their decision-making capability, LDCs will need to provide probabilistic daily demand forecasts. Quantiles can be fit to the errors produced by Saber's cross validation method using composite quantile regression. The biggest hurdle to generating probabilistic demand forecasts is to first obtain probabilistic weather forecasts.

Weather vendors have been unable to supply Marquette University GasDay™ with reliable probabilistic weather forecasts. However, much research in the meteorological community has been dedicated to probabilistic weather forecasting [8], lending confidence that they may become readily available. In the meantime, we propose a simple solution to modeling uncertainty of weather forecasts. First, a Laplacian distribution is assumed about the errors of weather forecasts. Then, the standard deviation and mean of the errors is tracked using a the same methods implemented in Brown's Dynamic Post Processor (described fully in [5]). The probabilistic weather forecast is derived by adding the point forecast to the Laplacian distribution.

BIBLIOGRAPHY

[1] L. Aguilera, W. Ramming, T. Monnig, and G. Becker, "2016 SGA gas forecasters survey results," Southern Gas Association, Tech. Rep., 2016.

[2] V. Barnett, *Environmental Statistics*. John Wiley and Sons., 2004.

[3] K. E. Broehl, "Gas peak design day analysis," Master's thesis, Wright State University, Dayton, OH, 7 1994.

[4] R. Brown, P. Kaefer, C. Jay, and S. Vitullo, "Forecasting natural gas design day demand from historical monthly data," *PSIG 2014 Conference Proceedings*, 5 2014.

[5] R. Brown, D. Kaftan, J. Smalley, M. Fakoor, S. Graupman, R. Povinelli, and G. F. Corliss, "Improving daily natural gas forecasting by tracking and combining models," in *37th Annual International Symposium on Forecasting*, 2017.

[6] 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 *Power & Energy Society General Meeting, 2015 IEEE*. IEEE, 2015, pp. 1–5.

[7] CBS Sports. (2018) Providence vs Marquette shot chart. [Online]. Available: https://www.cbssports.com/collegebasketball/gametracker/shotchart/NCAAB_20180103_MARQET@PROV [accessed 2018-04-03].

[8] A. Council, "Enhancing weather information with probability forecasts," *Bulletin of the American Meteorological Society*, vol. 89, pp. 1049–1053, 2008.

[9] A. C. Davison and R. L. Smith, "Models for exceedances over high thresholds," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 393–442, 1990.

[10] M. Debevc, "Design day demand forecast," *PSIG 2014 Conference Proceedings*, 5 2014.

[11] A. D'Silva, "Estimating the extreme low-temperature event using nonparametric methods," Master's thesis, Marquette University, Department of Electrical and Computer Engineering, Milwaukee, WI, 2015.

[12] Energy Information Administration, "U.S. energy facts." [Online]. Available: https://www.eia.gov/energyexplained/?page=us_energy_home [accessed 2018-04-03].

[13] M. Fréchet, "Sur la loi de probabilité de l'écart maximum," *Ann. de la Soc. polonaise de Math*, vol. 6:93, 1927.

[14] T. Gneiting, F. Balabdaoui, and A. Raftery, "Probabilistic forecasts, calibration and sharpness," *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, vol. 69, no. 2, pp. 243–268, 4 2007.

[15] Z. Gong, "Extreme value theory in periodic time series," Ph.D. dissertation, Clemson University, 2013.

[16] E. Gumbel, *Statistics of Extremes*. Echo Point Books and Media LLC., 1958.

[17] H. Hasan, N. Salam, and S. Kassim, "Modeling annual extreme temperature using generalized extreme value distribution: A case study in malaysia," *American Institute of Physics Conference Proceedings*, 2013.

[18] R. Hogg, personal communication with Koenker, R.

[19] 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, Jan 2014.

[20] T. Hong and S. Fan, "Probabilistic electric load forecasting: A tutorial review," *International Journal of Forecasting*, vol. 32, no. 3, pp. 914–938, 2016. [Online]. Available: http://EconPapers.repec.org/RePEc:eee:intfor:v:32:y:2016:i:3:p:914-938 [accessed 2018-04-03].

[21] J. R. M. Hosking and J. R. Wallis, "Parameter and quantile estimation for the generalized pareto distribution," *Technometrics*, vol. 29, no. 3, pp. 339 – 349, 1987. [Online]. Available: http://www.jstor.org/stable/1269343 [accessed 2018-04-03].

[22] W. K. Huang, M. L. Stein, D. J. McInerney, and E. J. Moyer, "Estimating changes in temperature extremes from millennial-scale climate simulations using generalized extreme value (GEV) distributions," *Advances in Statistical Climatology, Meteorology and Oceanography*, vol. 2, no. 1, p. 79, 2016.

[23] D. R. Hunter and K. Lange, "Quantile regression via an mm algorithm," *Journal of Computational and Graphical Statistics*, 2000.

[24] P. Kaefer, B. Ishola, R. H. Brown, and G. F. Corliss., "Using surrogate data to mitigate the risks of natural gas forecasting on unusual days." in *Proceedings of the 35th International Symposium on Forecasting*, 2015.

[25] D. Kaftan, A. D'Silva, G. Corliss, and R. Brown, "Determining extreme cold weather event conditions," in Progress.

[26] R. Koenker, *Quantile Regression*, ser. Econometric Society Monographs. Cambridge University Press, 2005.

[27] R. Koenker and J. Bassett, "Regression quantiles," *Econometrica*, vol. 46, no. 1, pp. 33–50, 1978. [Online]. Available: http://www.jstor.org/stable/1913643 [accessed 2018-04-03].

[28] R. Koenker and B. J. Park, "An interior point algorithm for nonlinear quantile regression," *Journal of Econometrics*, vol. 71, no. 1, pp. 265 – 283, 1996. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0304407696845076 [accessed 2018-04-03].

[29] R. W. Koenker and V. D'Orey, "Algorithm as 229: Computing regression quantiles," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 36, no. 3, pp. 383–393, 1987. [Online]. Available: http://www.jstor.org/stable/2347802 [accessed 2018-04-03].

[30] S. Kotz and S. Nadarajah, *Extreme Value Distributions: Theory and Applications*. World Scientific, 2000.

[31] F. K. Lyness, "Consistent forecasting of severe winter gas demand," *The Journal of the Operational Research Society*, vol. 32, no. 5, pp. 347–459, 1981.

[32] G. Merkel, "Deep neural networks as time series forecasters of energy demand," Master's thesis, Marquette University, 2017.

[33] I. Miller and M. Miller, *John E. Freund's mathematical statistics with applications*. Pearson, 2015.

[34] J. Nowotarski and R. Weron, "Recent advances in electricity price forecasting: A review of probabilistic forecasting," *Renewable and Sustainable Energy Reviews*, vol. 81, pp. 1548–1568, 2017.

[35] A. Papoulis and S. U. Pillai, *Probability, Random Variables and Stochastic Processes*, 4th ed. Boston, MA: McGraw-Hill, 2002.

[36] P. Pinson, H. Nielsen, J. Moller, H. Madsen, and G. Kariniotakis, "Non-parametric probabilistic forecasts of wind power: Required properties and evaluation," *Wind Energy*, vol. 10, p. 497516, 2007.

[37] H. Pishro-Nik, *Introduction to probability, statistics, and random processes*. Kappa Research, LLC, 2016.

[38] M. Saber, "Quantifying forecast uncertainty in the energy domain," Ph.D. dissertation, Marquette University, 2017.

[39] C. Scarrott and A. MacDonald, "A review of extreme value threshold estimation and uncertainty quantification," *REVSTAT–Statistical Journal*, vol. 10, no. 1, pp. 33–60, 2012.

[40] B. Seaman, "Retail sales forecasting at Walmart," in *Proceedings of the 37th International Symposium on Forecasting*, 2017.

[41] J. Smalley, B. Ishola, G. Corliss, and R. Brown, "Prior day weather sensitivity in natural gas demand," *In Process*, 2017.

[42] S. Tenneti, "Identification of nontemperature-sensitive natural gas customers and forecasting their demand," Master's thesis, Marquette University, Department of Electrical and Computer Engineering, Milwaukee, WI, May 2009.

[43] U.S. Energy Information Administration, "Annual energy outlook 2017," 2017. [Online]. Available: https://www.eia.gov/outlooks/aeo/pdf/0383(2017).pdf [accessed 2018-04-03].

[44] S. R. Vitullo, R. H. Brown, G. F. Corliss, and B. M. Marx, "Mathematical models for natural gas forecasting," *Canadian Applied Mathematics Quarterly*, vol. 17, no. 4, pp. 807–827, Jan. 2009.

[45] H. Von Storch, "Misuses of statistical analysis in climate research," in *Analysis of Climate Variability*. Springer, 1999, pp. 11–26.

[46] H. J. Wang, D. Li, and X. He, "Estimation of high conditional quantiles for heavy-tailed distributions," *Journal of the American Statistical Association*, vol. 107, no. 500, pp. 1453–1464, 2012.

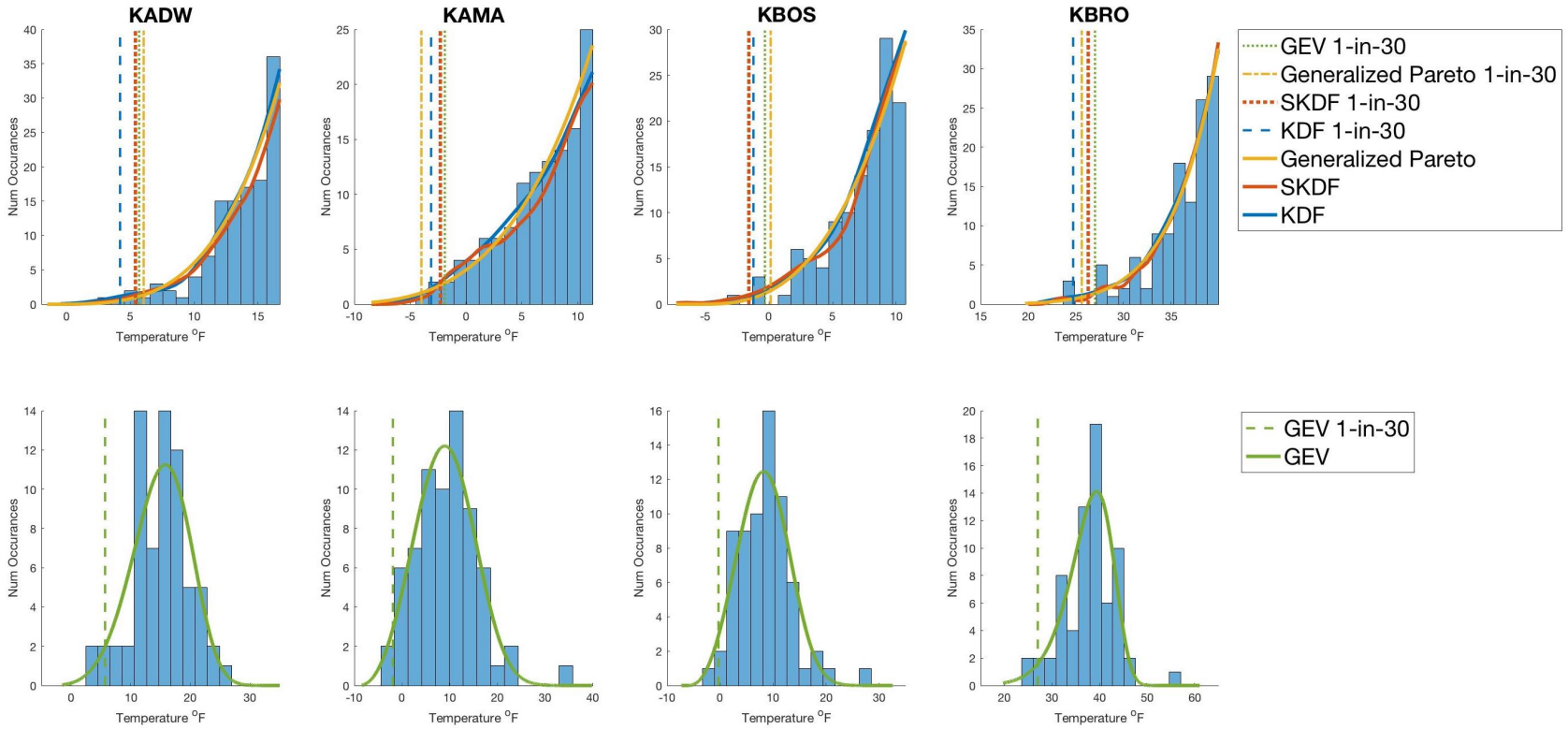
[47] K. Wang and H. J. Wang, "Optimally combined estimation for tail quantile regression," *Statistica Sinica*, pp. 295–311, 2016.

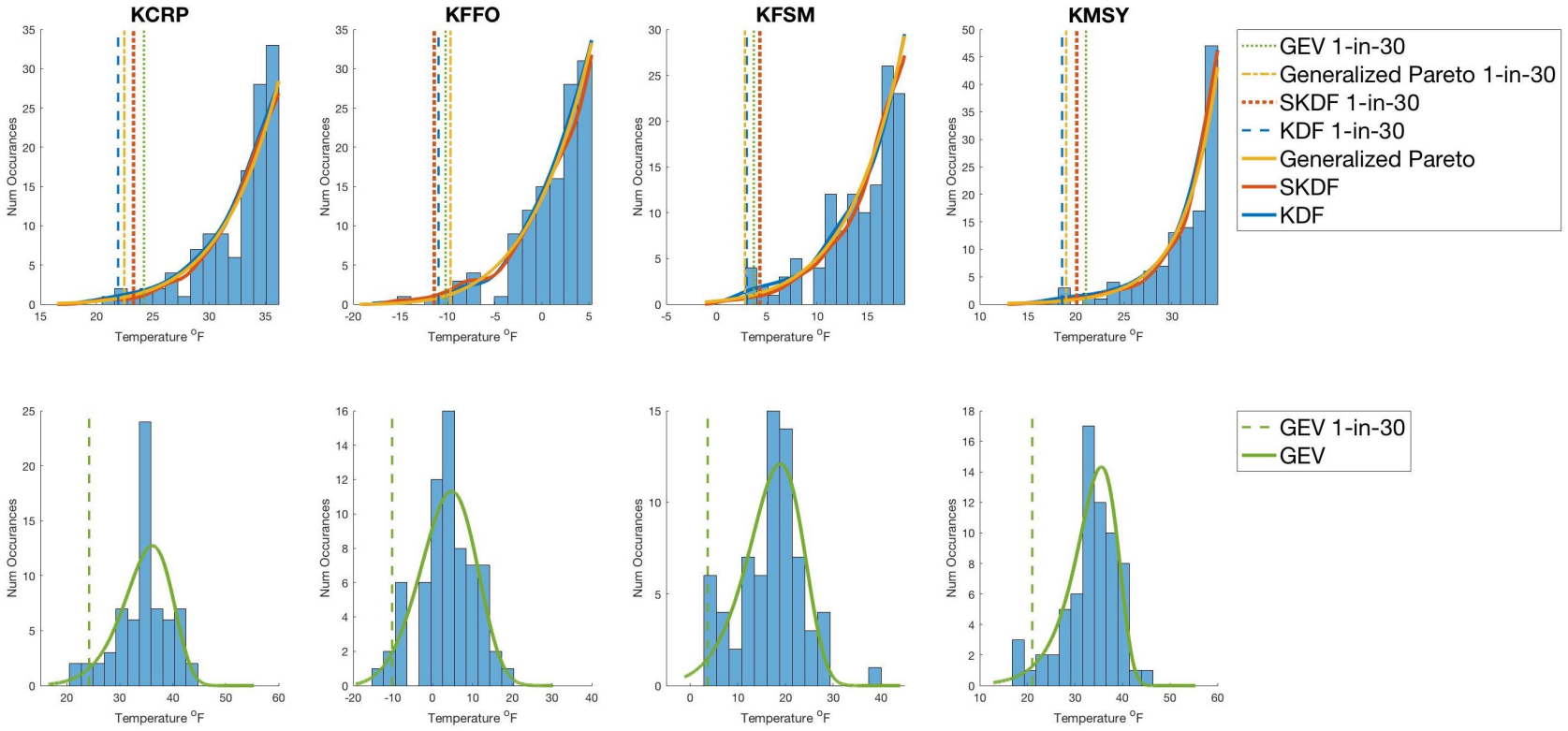
[48] S. Zheng, "Gradient descent algorithms for quantile regression with smooth approximation," *International Journal of Machine Learning and Cybernetics*, 2011.

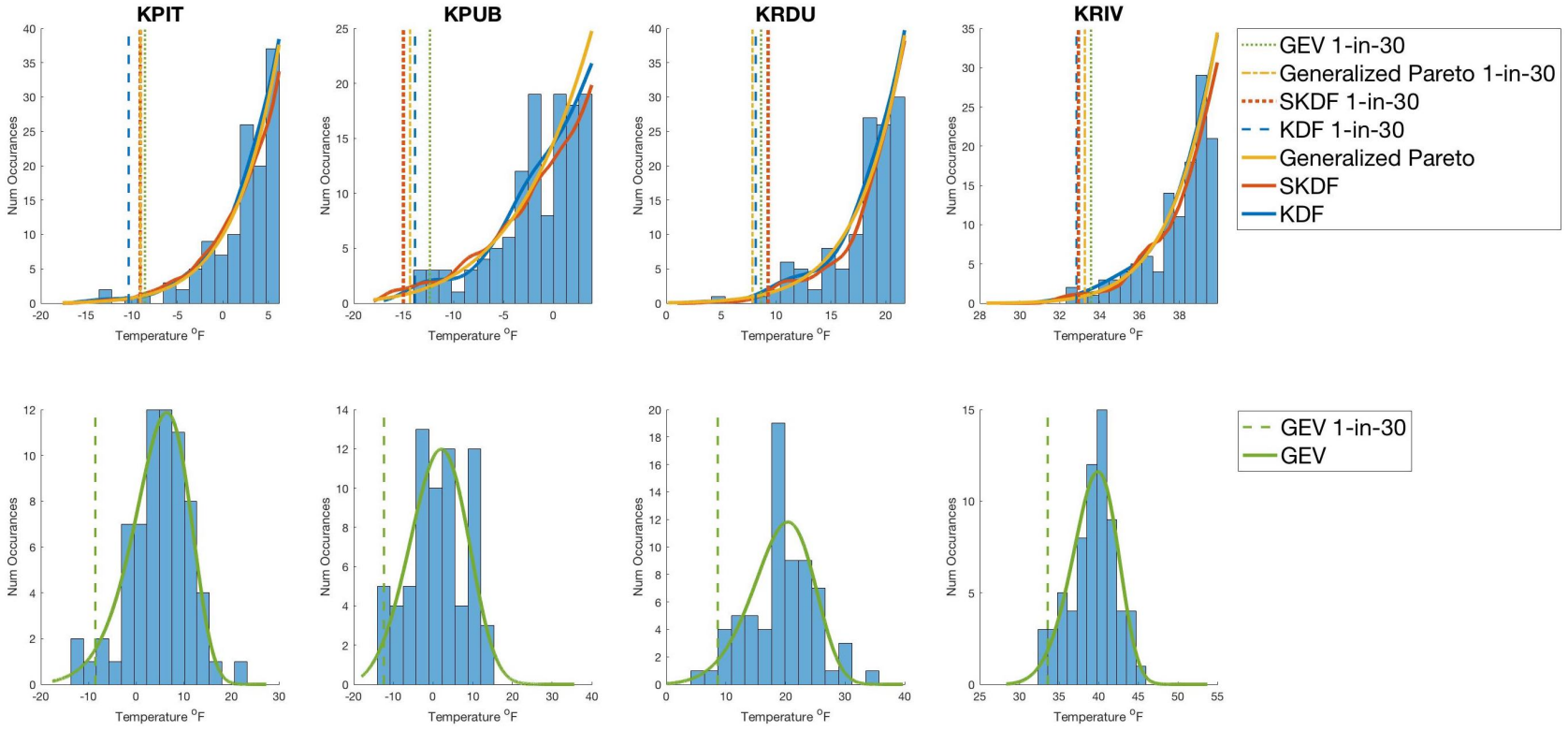
Appendix A

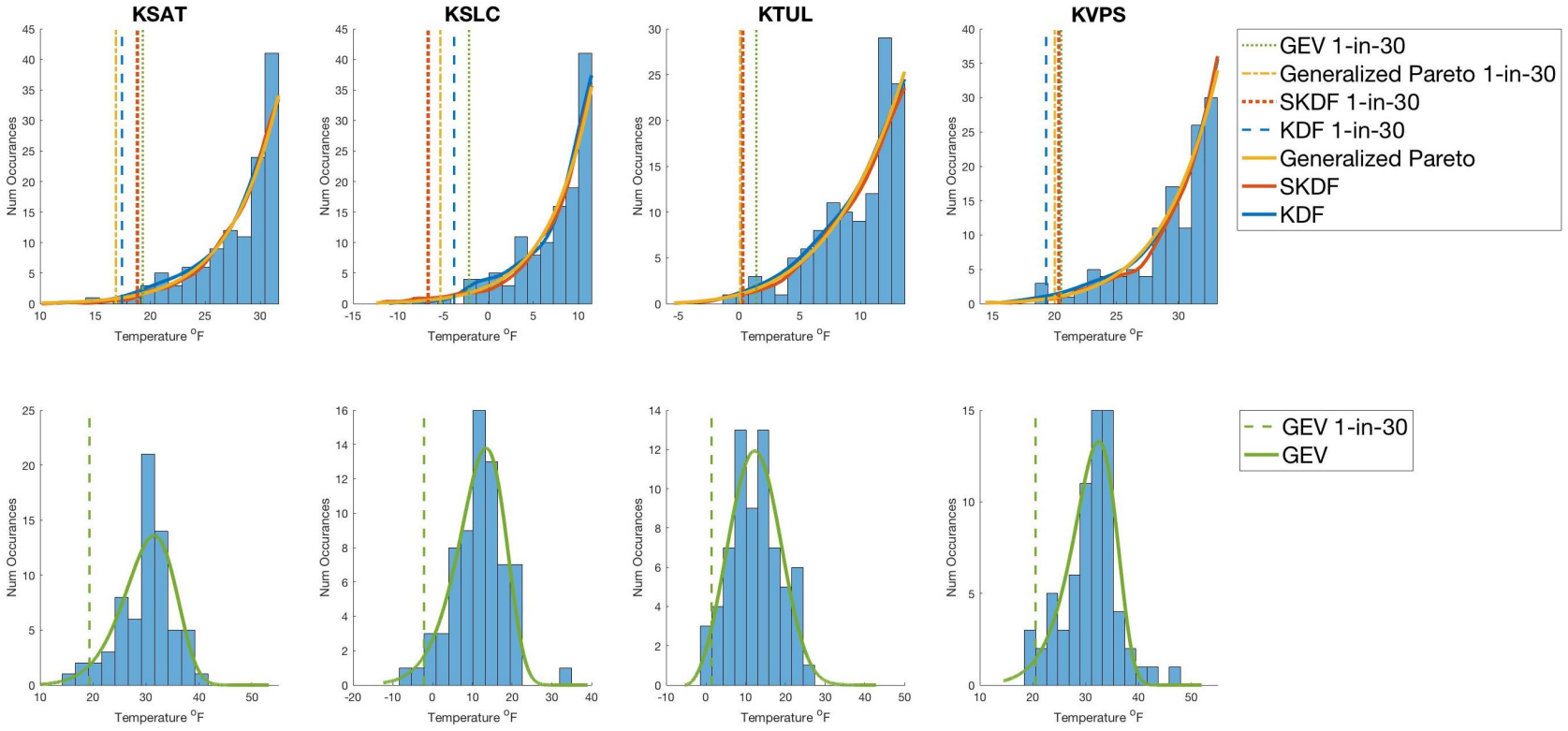
Raw Temperature Distribution Fit For All Stations

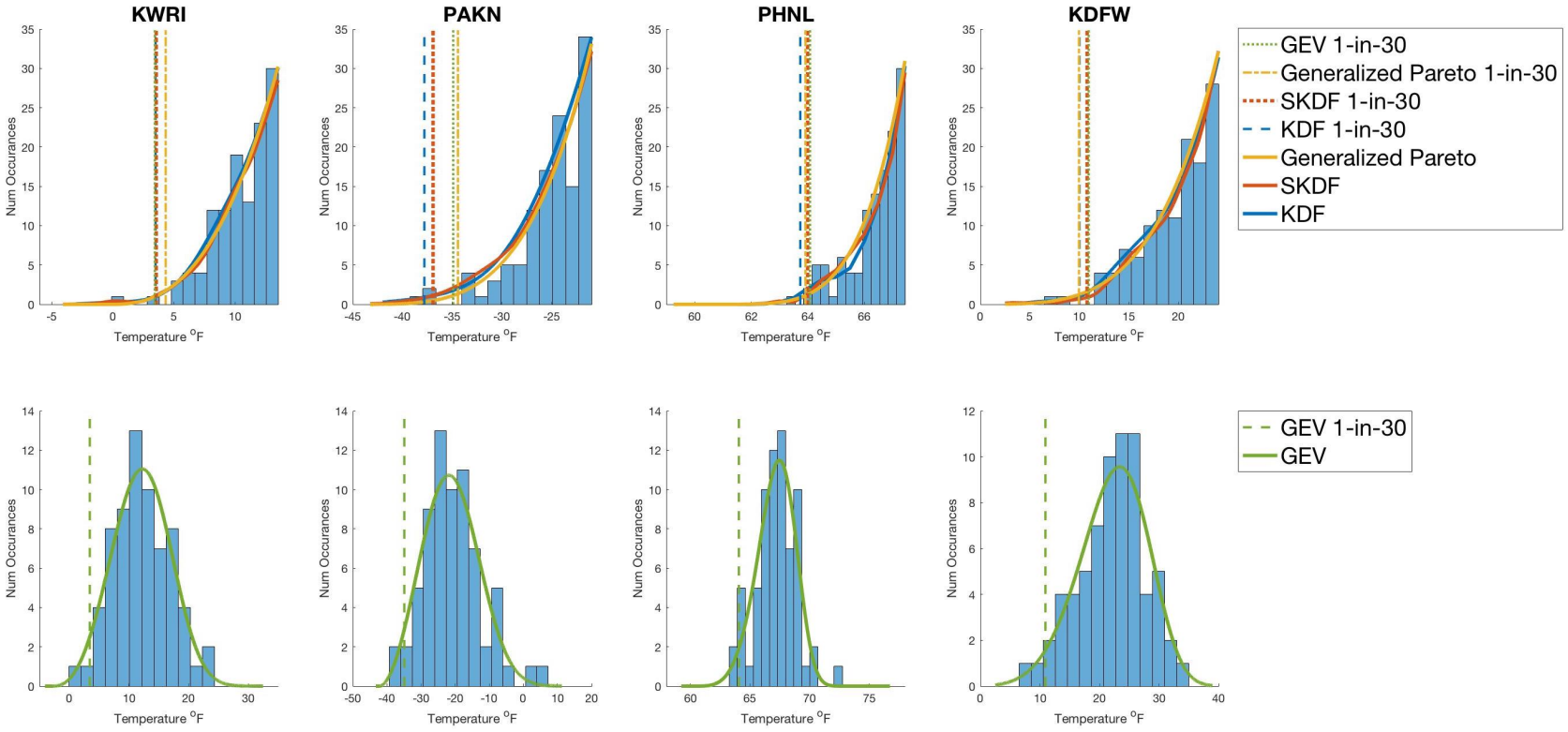
It would have been too cumbersome to visualize the distributions fit to each of the 38 stations in the *continental* data set. However, for the sake of transparency, we include them here.

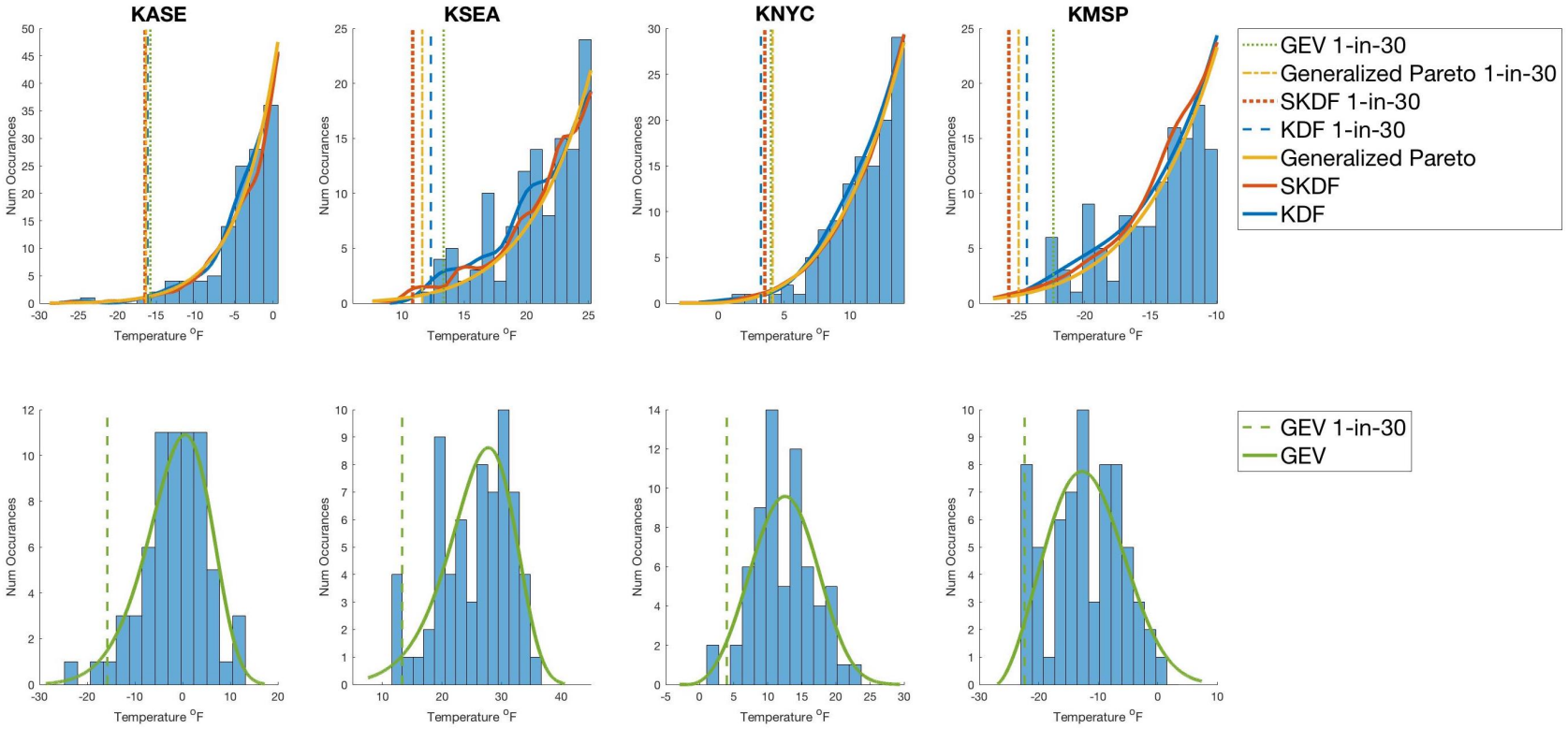

KADW
KAMA
KBOS
KBRO
Num Occurances
Temperature °F
GEV 1-in-30
Generalized Pareto 1-in-30
SKDF 1-in-30
KDF 1-in-30
Generalized Pareto
SKDF
KDF
GEV 1-in-30
GEV

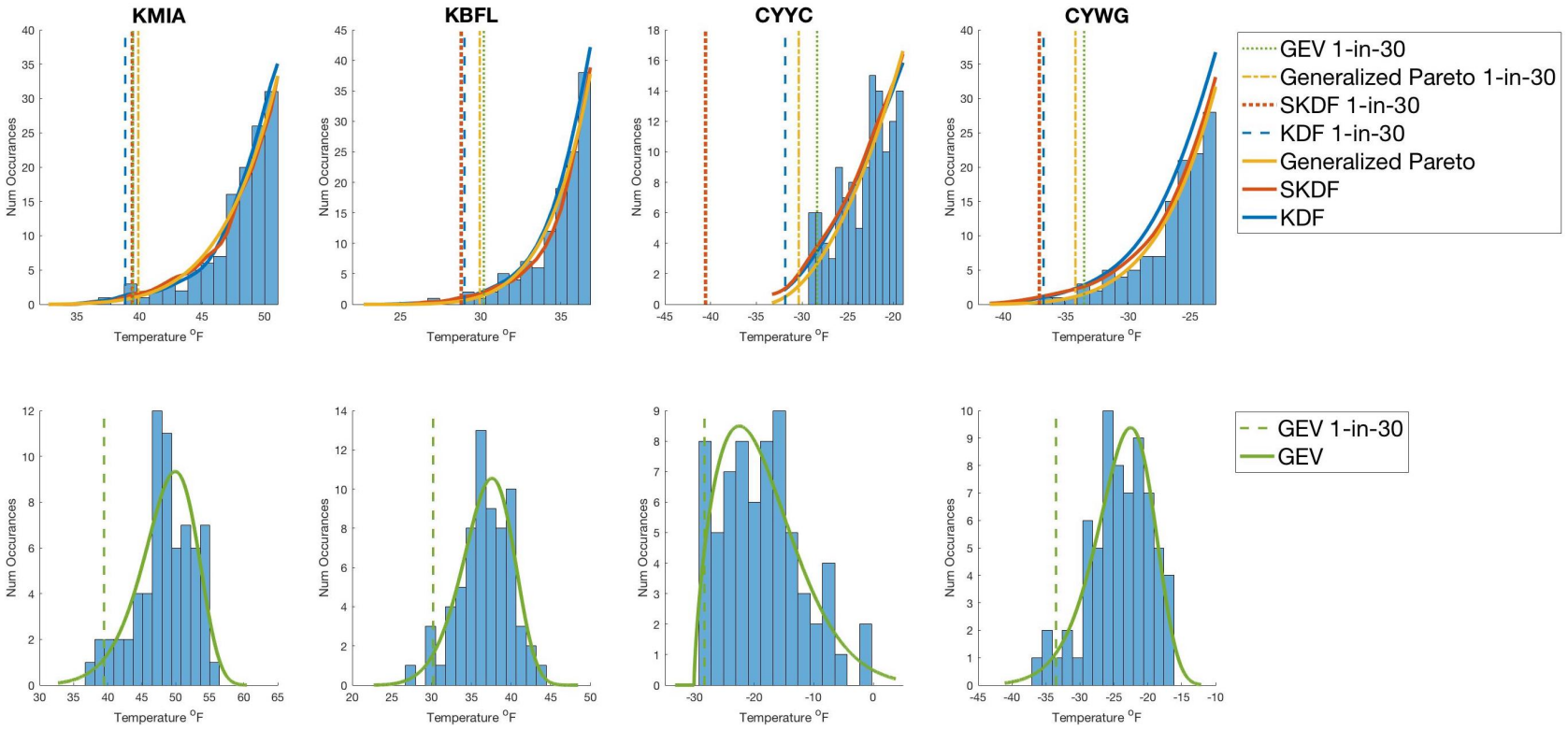

KCRP
KFFO
KFSM
KMSY
Num Occurances
Temperature °F
GEV 1-in-30
Generalized Pareto 1-in-30
SKDF 1-in-30
KDF 1-in-30
Generalized Pareto
SKDF
KDF
GEV 1-in-30
GEV

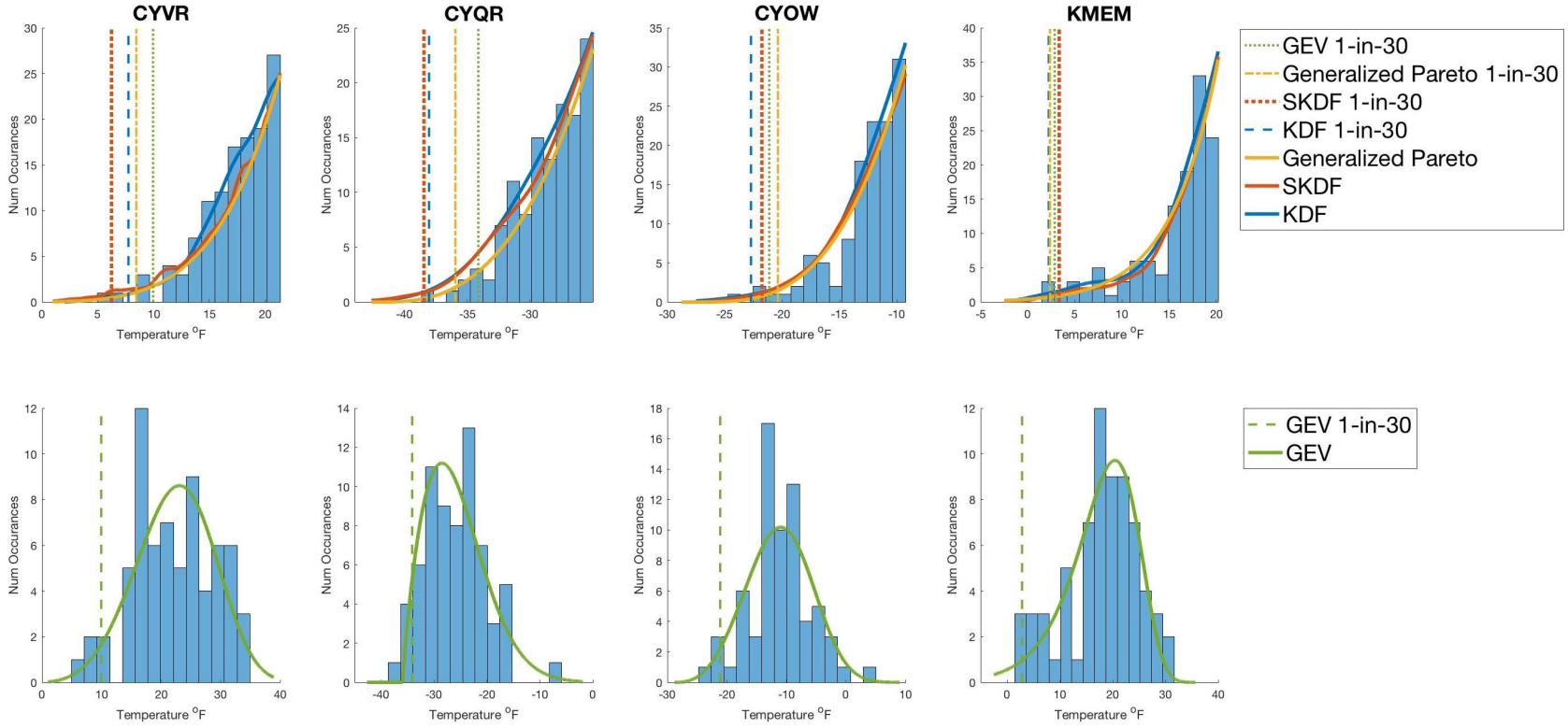

KPIT
KPUB
KRDU
KRIV
Num Occurances
Temperature °F
GEV 1-in-30
Generalized Pareto 1-in-30
SKDF 1-in-30
KDF 1-in-30
Generalized Pareto
SKDF
KDF
GEV 1-in-30
GEV

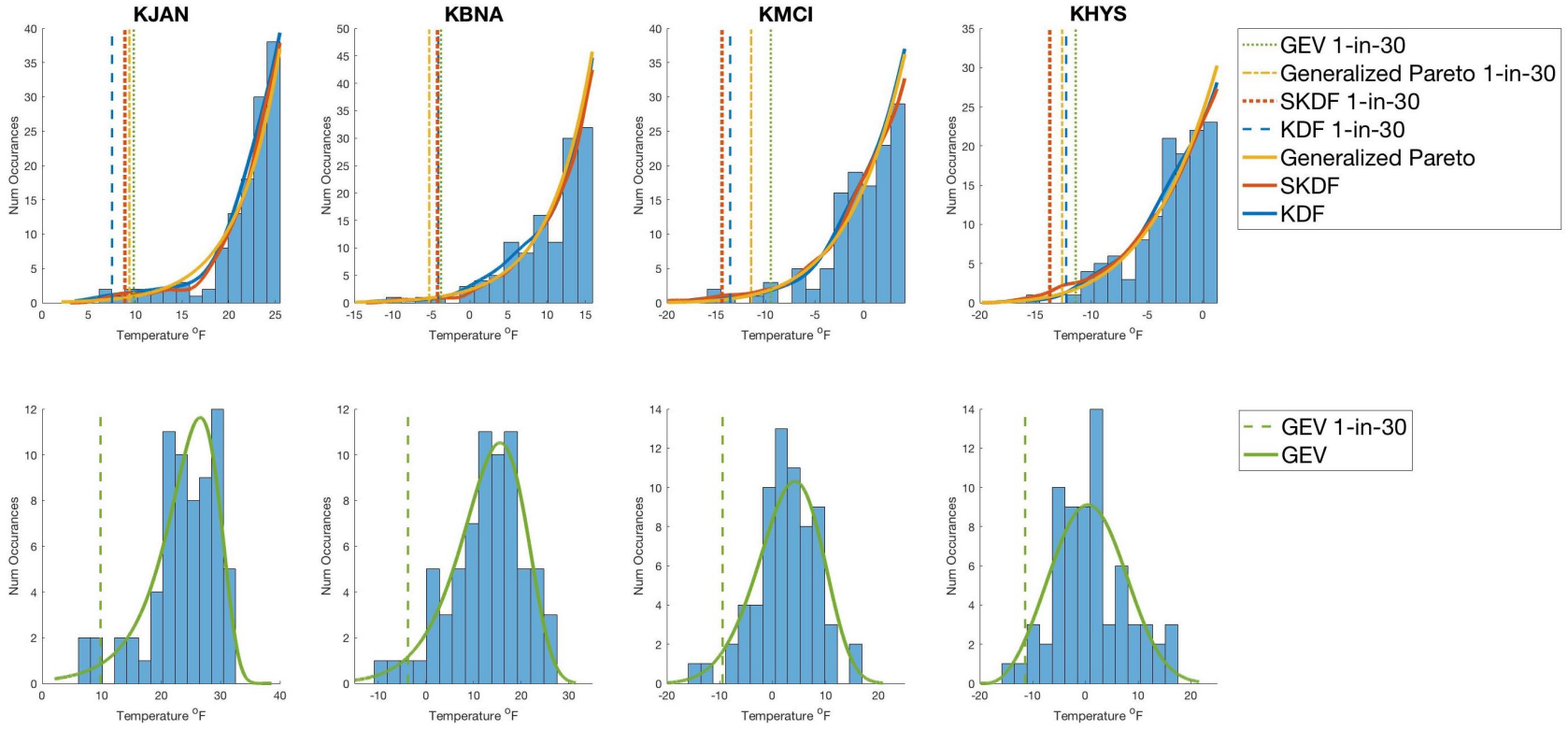

KSAT
KSLC
KTUL
KVPS
Num Occurances
Temperature °F
GEV 1-in-30
Generalized Pareto 1-in-30
SKDF 1-in-30
KDF 1-in-30
Generalized Pareto
SKDF
KDF
GEV 1-in-30
GEV

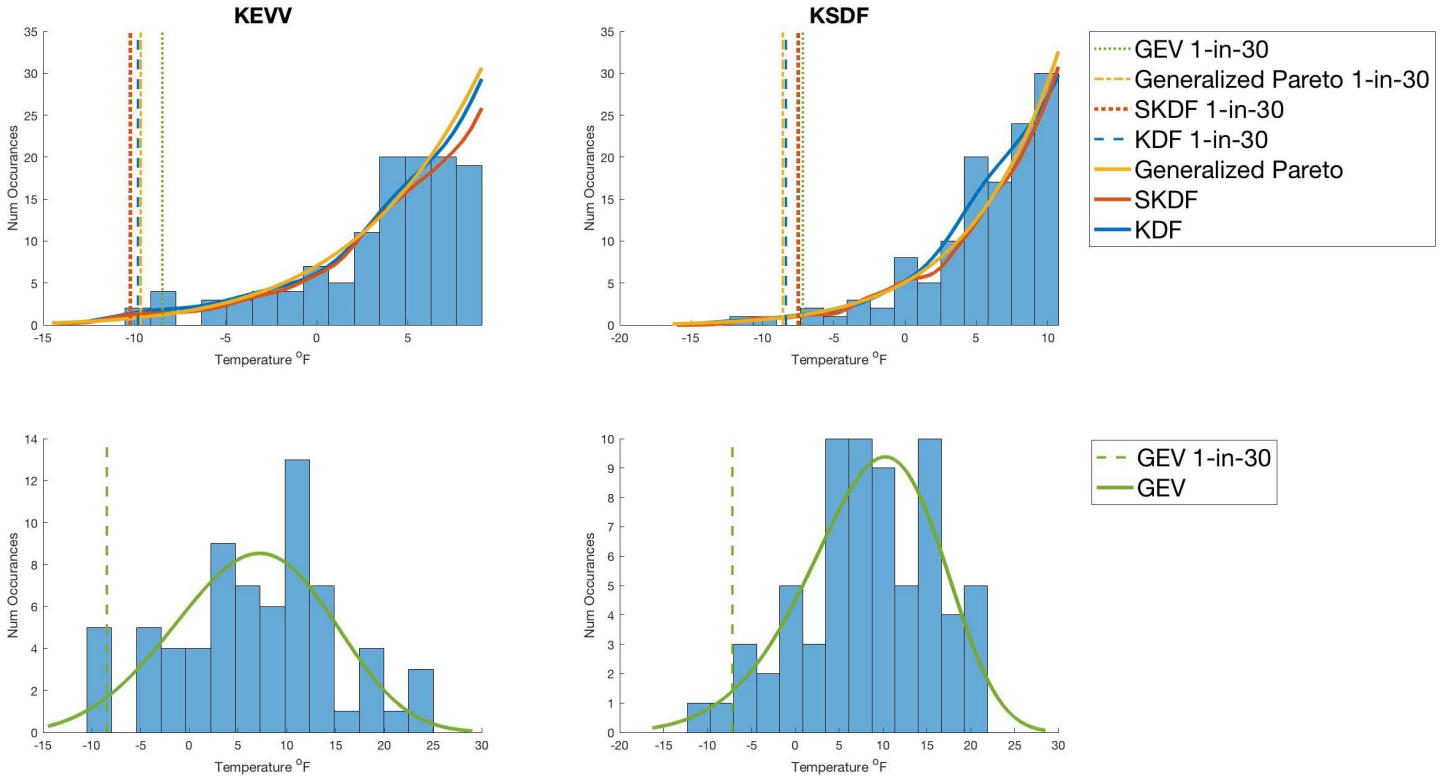

KWRI
PAKN
PHNL
KDFW
Num Occurances
Temperature °F
GEV 1-in-30
Generalized Pareto 1-in-30
SKDF 1-in-30
KDF 1-in-30
Generalized Pareto
SKDF
KDF
GEV 1-in-30
GEV


KASE
KSEA
KNYC
KMSP
Num Occurances
Temperature °F
GEV 1-in-30
Generalized Pareto 1-in-30
SKDF 1-in-30
KDF 1-in-30
Generalized Pareto
SKDF
KDF
GEV 1-in-30
GEV


KMIA
KBFL
CYYC
CYWG
Num Occurances
Temperature °F
GEV 1-in-30
Generalized Pareto 1-in-30
SKDF 1-in-30
KDF 1-in-30
Generalized Pareto
SKDF
KDF
GEV 1-in-30
GEV


CYVR
CYQR
CYOW
KMEM
Num Occurances
Temperature °F
GEV 1-in-30
Generalized Pareto 1-in-30
SKDF 1-in-30
KDF 1-in-30
Generalized Pareto
SKDF
KDF
GEV 1-in-30
GEV


KJAN
KBNA
KMCI
KHYS
Num Occurances
Temperature °F
GEV 1-in-30
Generalized Pareto 1-in-30
SKDF 1-in-30
KDF 1-in-30
Generalized Pareto
SKDF
KDF
GEV 1-in-30
GEV


KEVV
KSDF
Num Occurances
Temperature °F
GEV 1-in-30
Generalized Pareto 1-in-30
SKDF 1-in-30
KDF 1-in-30
Generalized Pareto
SKDF
KDF
GEV 1-in-30
GEV