Marquette University
e-Publications@Marquette

Master's Theses (2009 -) | Dissertations, Theses, and Professional Projects

Estimating the Extreme Low-Temperature Event using Nonparametric Methods

Anisha D'Silva
Marquette University

Recommended Citation
D'Silva, Anisha, "Estimating the Extreme Low-Temperature Event using Nonparametric Methods" (2015). *Master's Theses (2009 -)*. Paper 292.
http://epublications.marquette.edu/theses_open/292

ESTIMATING THE EXTREME LOW-TEMPERATURE EVENT USING NONPARAMETRIC METHODS

by

Anisha D'Silva,
Bachelor of Science in Biomedical Engineering
(Major in Bioelectronics, Minor in Mathematics, Minor in Biology)

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 2015

ABSTRACT

ESTIMATING THE EXTREME LOW-TEMPERATURE EVENT
USING NONPARAMETRIC METHODS

Anisha D'Silva,
Bachelor of Science in Biomedical Engineering
(Major in Bioelectronics, Minor in Mathematics, Minor in Biology)

Marquette University, 2015

This thesis presents a new method of estimating the one-in-N low temperature threshold using a non-parametric statistical method called kernel density estimation applied to daily average wind-adjusted temperatures. We apply our One-in-N Algorithm to local gas distribution companies (LDCs), as they have to forecast the daily natural gas needs of their consumers. In winter, demand for natural gas is high. Extreme low temperature events are not directly related to an LDCs gas demand forecasting, but knowledge of extreme low temperatures is important to ensure that an LDC has enough capacity to meet customer demands when extreme low temperatures are experienced.

We present a detailed explanation of our One-in-N Algorithm and compare it to the methods using the generalized extreme value distribution, the normal distribution, and the variance-weighted composite distribution. We show that our One-in-N Algorithm estimates the one-in-N low temperature threshold more accurately than the methods using the generalized extreme value distribution, the normal distribution, and the variance-weighted composite distribution according to root mean square error (RMSE) measure at a 5% level of significance. The One-in-N Algorithm is tested by counting the number of times the daily average wind-adjusted temperature is less than or equal to the one-in-N low temperature threshold.

ACKNOWLEDGMENTS

Anisha D'Silva,
Bachelor of Science in Biomedical Engineering
(Major in Bioelectronics, Minor in Mathematics, Minor in Biology)

Writing this thesis to achieve a Master degree in Electrical Engineering has been an incredible journey for me. Like other aspiring advanced degree students, I have experienced several ups and downs in this endeavor to complete the research and writing for this thesis, and I am grateful for the support of friends.

As I near the end of this journey, I take a step back and remember those first moments when I first started working in the GasDay lab. I was excited to be part of this project, to learn about natural gas distribution, and to make a contribution to this project. Today, I feel that the entire GasDay team is my family. Everyone supported my research, provided me with constructive criticism, and most importantly, provided me with genuine words of encouragement when things looked gloomy and I was not sure that I could complete this project. Thank you!

Particularly, I would like to express my gratitude to my thesis advisors, Dr. Ronald Brown and Dr. George Corliss for giving me the opportunity to work for GasDay and for all the help, motivation, and financial support they provided, without which none of this would be possible. They have both been my mentors in both technical and personal matters, and I am grateful to have such great role models in my life. I am thankful to my committee members, Dr. Susan Schneider and Dr. Monica Adya, for always encouraging me to complete this project and have spent a lot of time in the discussion of the content of this thesis.

I would like to thank Steve Vitullo, Tsugi Sakauchi, Hermine Akouemo, Bo Pang, and Yifan Li, for technical help and sharing ideas during the course of this thesis. We bonded over seminar presentations and classes that we took together. Also, I would like to thank Paula Gallitz for all her support, words of encouragement, and for helping me network with others. I am thankful to John-Paul Kucera and Dennis Wiebe, for their support and encouragement as I worked to complete this thesis.

Finally, I would like to thank my family, Ryan Erickson, Nikhil D'Silva, Amy Soares, Clement Soares, Alice D'Silva, and Wilfred D'Silva, for always being my rock, both financially and emotionally. Without their support, it would be difficult to complete this project.

TABLE OF CONTENTS

ACKNOWLEDGMENTS . . . . . **i**

LIST OF TABLES . . . . . **iv**

LIST OF FIGURES . . . . . **v**

CHAPTER 1 Estimating Extreme Low-Temperature Weather Events **1**

1.1 Natural Gas 2

1.2 Problem Statement 4

1.3 Proposed Solution 7

1.4 Thesis Outline 10

CHAPTER 2 Literature Survey on the Likelihood of Rare Events and Statistical Theory **13**

2.1 Survey of Rare Events in Meteorology, Ecology, and the Nuclear Power Industry 13

2.2 Parametric and Nonparametric Statistical Theory 32

2.2.1 Probability Density Functions 32

2.2.2 Cumulative Distribution Functions 33

2.2.3 Gaussian (Normal) Distribution 36

2.2.4 Log Normal Distribution 36

2.2.5 Generalized Extreme Value Distribution 41

2.2.6 Kernel Density Estimation 44

2.2.7 Variance-weighted composite distribution 46

CHAPTER 3 Estimating the One-in-*N* Coldest Temperature Threshold **51**

3.1 Summary of Problem 51

3.2 Preparing the Data 54
3.3 Output of the One-in-N Algorithm 63
3.4 The One-in-N Algorithm 65
CHAPTER 4 One-in-N Algorithm Test Results and Discussion of Output 72
4.1 Overview 72
4.2 Analysis of the One-in-N Algorithm Output 73
4.2.1 Discussion: Extreme Cold Threshold for KMKE in Milwaukee 76
4.2.2 Discussion: Extreme Cold Threshold for KABQ in Albuquerque 79
4.2.3 Discussion: Extreme Cold Threshold for PANC in Anchorage 83
4.3 Test and Results 86
CHAPTER 5 Conclusions and Future Research 90
5.1 Conclusions 90
5.2 Future Research 91
5.2.1 Extension of Work 92
5.2.2 Hypotheses to be Explored 94
BIBLIOGRAPHY 97

LIST OF TABLES

2.1 Kernel Functions 45

3.1 Common notation used in Chapter 3 52

3.2 Size of data set 57

3.3 "Winter" for each weather station assuming non-leap years 63

3.4 One-in-N threshold temperature (°F) using the one-in-N algorithm . . . 70

4.1 Validation test results after 100 iterations 88

LIST OF FIGURES

1.1 Natural gas industry [1] 3

1.2 Daily temperature for typical winter days for every year in the data set . 9

2.1 Schematic diagram depicting how a change in mean can affect extreme events [2] 15

2.2 Schematic diagram depicting how a change in variance can affect extreme events [2] 15

2.3 Schematic diagram depicting how a change in mean and variance can affect extreme events [2] 16

2.4 Example of a probability density function 34

2.5 Example of a cumulative distribution function 35

2.6 Gaussian (normal) probability density function 37

2.7 Gaussian (normal) cumulative distribution function 38

2.8 Log normal probability density function 39

2.9 Log normal cumulative distribution function 40

2.10 Three types of the generalized extreme value probability density function 42

2.11 Three types of the generalized extreme value cumulative distribution function 43

2.12 Cartoon of Construction of the Kernel Density Function [3] 46

2.13 Comparison between the Gamma probability density function and its Kernel Estimate 47

2.14 Comparison between the Gamma cumulative distribution function and its Kernel Estimate 48

3.1 Steps for the one-in-N algorithm 55

3.2 Daily average wind-adjusted temperature for Milwaukee, WI (KMKE) . 57

3.3 Daily average wind-adjusted detrended temperature for Milwaukee, WI . 60

3.4 KMKE coldest 91 days × n years temperature histogram 65

3.5 KMKE coldest 91 days × n years temperature with distributions 66

4.1 Increasing Window - Milwaukee, WI (KMKE) conditions by years of data to determine the minimum number of years needed to calculate the low temperature threshold 75

4.2 Sliding Window - Milwaukee, WI (KMKE) conditions using a window of 20 years of data to determine if conditions change with time 78

4.3 Increasing Window - Albuquerque, NM (KABQ) conditions by years of data to determine the minimum number of years needed to calculate the low temperature threshold 80

4.4 Sliding Window - Albuquerque, NM (KABQ) conditions using a window of 20 years of data to determine if conditions change with time 82

4.5 Increasing Window - Anchorage, AK (PANC) conditions by years of data to determine the minimum number of years needed to calculate the low temperature threshold 84

4.6 Sliding Window - Anchorage, AK (PANC) conditions using a window of 20 years of data to determine if conditions change with time 85

CHAPTER 1

Estimating Extreme Low-Temperature Weather Events

How cold is it going to be this winter? Weather has always been a topic of interest because it affects daily life. People need to know what the weather forecast is for the present day, the next day, the next week, and for other intervals of time. Knowledge of weather conditions helps to plan some common activities such as travel, everyday commuting to work or school, and budgeting energy resources. Since weather is such an integral part of people's lives, extremes in weather can have serious and possibly devastating consequences to society, infrastructure, and animal life. Hence, extreme weather events receive much attention in news reports on climate [2]. Tornadoes, hurricanes, drought, floods, extreme high-temperature events, and extreme low-temperature events are examples of extreme weather conditions.

A big challenge associated with extreme weather is to minimize its impact on daily life. To meet this challenge, one needs to know if there is a pattern that determines when these events occur and how severe they may be. Statistics help in finding this pattern and thus help estimate the probability of having an extreme

event. With this knowledge, people can take precautions to reduce the negative impact of extreme weather events.

This thesis presents a procedure to estimate the threshold temperature defining an extreme low-temperature event that occurs, on average, once in N years. Our work is motivated by the desire to forecast natural gas demands accurately during an extreme low-temperature event. Although this thesis work does not directly help forecast gas consumption for a particular temperature, it does help forecast the maximum consumption that the local gas distribution company (LDC) is likely to experience. In the next section, we will discuss the natural gas industry briefly.

1.1 Natural Gas

According to Potocnik [1], natural gas is a naturally occurring combustible mixture of gaseous hydrocarbons in reservoirs of porous rock capped by impervious strata. Like petroleum, it is formed during the decomposition of organic matter in sedimentary deposits, and it is a non-renewable resource that takes millions of years to form. It consists largely of methane (CH_4). Natural gas is one of the cleanest sources of energy, and it accounts for roughly 25% of the total energy consumption in the United States. It is used to heat homes, to cook food, to generate electricity, and for other domestic and industrial uses. In the United States, natural gas is

transported via pipelines and increasingly in the form of liquefied natural gas (LNG) through tanks. Figure 1.1 describes a simplified process of how natural gas is transported from well-head to the end user.

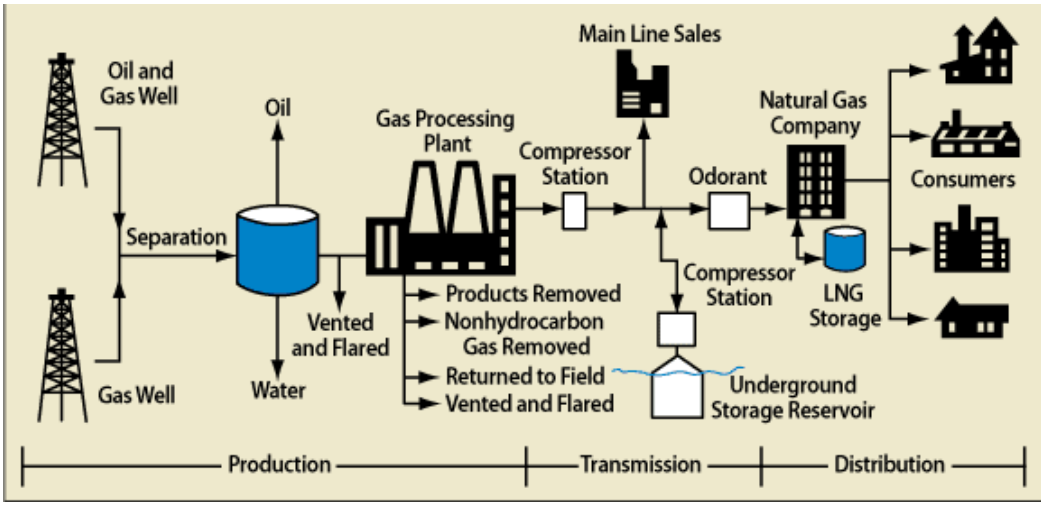


Figure 1.1: Natural gas industry [1]

In this work, we are concerned with the distribution section of the natural gas industry, seen in Figure 1.1. A local gas distribution company (LDC) must ensure that it meets the daily gas requirements of its consumers. To achieve this goal successfully and to ensure customer satisfaction, LDCs have to forecast the daily natural gas needs of their consumers. Forecasting daily natural gas consumption accurately can be challenging. In winter, demand for natural gas increases because natural gas is largely used to heat space in homes and businesses,

while in summer, the demand for natural gas decreases as consumers use it mostly for the other domestic and industrial purposes including electric power generation [4]. Low temperature thresholds are not directly related to an LDCs gas demand forecasting, but knowledge of extreme low temperatures is important to ensure that an LDC has enough capacity to meet customer demands when extreme low temperatures are experienced.

In February 2011, New Mexico, Arizona, and Texas experienced unusually cold and windy weather, which resulted in natural gas production declines [5]. These declines ultimately resulted in natural gas curtailments or outages to more than 50,000 customers in these three states. The Federal Energy Regulatory Commission (FERC) initiated an inquiry into the Southwest outages and service disruptions to explore solutions that can mitigate future natural gas outages.

1.2 Problem Statement

In this thesis, we estimate the threshold defining an extreme cold temperature event that may be expected to occur, on average, once in N years for different weather stations in the United States using a non-parametric statistical method called kernel density estimation.

A local gas distribution company (LDC) is responsible for delivering natural

gas to its customers daily. Since natural gas is difficult to store, there is a demand for accurate forecasting models. If an LDC purchases too much or too little natural gas, there are high costs associated with these errors in forecasting gas demand. Extreme low-temperature events are not directly related to a utility's daily gas demand forecasting, but extreme low-temperature data is important for infrastructure capacity planning and for supply planning to ensure that the utility has sufficient capacity to supply gas to its customers during an extreme low-temperature event. We estimate the threshold defining an extreme cold temperature event that may be expected to occur, on average, once in N years for different weather stations in the United States, where N can be determined by a gas utility depending on the rarity of the cold event they need to analyze. In this thesis, we assume stationarity of weather and climate. We also assume that daily temperature is independent of neighboring days, but not identically distributed.

Let X_t be the random variable describing the average daily wind-adjusted temperature on a day, where the domain of the random variable is the set of all days in the entire historical record, and the range is the set of all possible temperatures. For our experiment, we use n years, then t is the index of days $\{1, 2, 3, ..., n \times 365\}$. We estimate a threshold temperature, T_{th}, with the property that the event $X_t \leq T_{th}$, may be expected to occur, on average, once in N years. Henceforth, we will refer to this threshold, T_{th}, as "one-in-N low temperature

threshold." We will consider $N = 0.25$, 0.5, 1, 2, 5, 10, 20, and 30. Below, we will discuss the properties of the one-in-N low temperature threshold.

We define an indicator function, $f_c(X_t)$, whose domain is temperature (X_t), and the range is the set $\{0, 1\}$. The rule is

$$f_c(X_t) = \begin{cases} 1 & \text{if } X_t \leq T_{th} \ , \\ \\ 0 & \text{otherwise} \ . \end{cases} \tag{1.1}$$

Our experiment is to count the number of times the temperature (X_t) falls below the threshold temperature (T_{th}). Our outcome is the independent random variable, *count*, which is the count of the number of extreme low temperature events in the n-year period.

$$count = \sum_{t=1}^{n \times 365} f_c(X_t) \ . \tag{1.2}$$

Hence, if we perform several experiments with different sets of n years of data, the expected value of the *count* should be

$$E\left(count\right) = E\left(\sum_{t=1}^{n \times 365} f_c(X_t)\right) = \frac{n}{N} \ . \tag{1.3}$$

For example, if we had $n = 300$ years and $N = 30$ years, then if we perform several experiments of counting the number of times the temperature falls below the

one-in-N low temperature threshold, each with 300 years of data, we should get $E\,(count) = 10$.

We have identified and described the problem of estimating the one-in-N low temperature threshold in this section and explained the property of the one-in-N low temperature threshold. We now introduce our proposed solution.

1.3 Proposed Solution

We organize this section starting with a brief overview of the data used for this thesis, followed by identification of the current methods, and a brief description of the proposed solution. At the end of this section, the reader will appreciate the proposed solution compared to the current methods.

The data set considered in this thesis consists of 264 weather stations associated with Marquette University's partner LDCs across the United States. Analysis and tests will be performed on weather stations with more than 30 years of data because of the values of $N = 0.25$, 0.5, 1, 2, 5, 10, 20, and 30 selected for experimentation in this thesis. For each year, we consider only typical winter data for our evaluation, which we explain in Chapter 3. In this thesis, we will discuss the results for only three weather stations chosen for their unique weather patterns as explained in Chapter 4. These weather stations are Milwaukee, WI (KMKE);

Albuquerque, NM (KABQ); and Anchorage, AK (PANC). KMKE depicts a typical weather pattern, while the other two weather stations reveal unexpected weather patterns.

Currently, probability density functions for several distributions are being used to fit the winter weather data to find the one-in-N year low temperature threshold. They are the normal, Weibull, Gumbel, generalized extreme value, logistic, Student t-location-scale, and a distribution created by a weighted variance of the aforementioned distributions. In Chapter 2, we will explain some of these distributions theoretically and discuss the results from the survey of literature conducted in research areas dealing with extreme rare conditions.

Since we are interested in the low temperature threshold, we would like to find a cumulative density function that best fits the left tail of the winter data. Some of the currently used distributions fit the winter data better than some of the other distributions. For a visual understanding of how well a distribution fits the data, we have fit the probability density function of four distributions and compared them to the histogram of temperature data, as shown in Figure 1.2. The goodness of fit is measured using an error score (RMSE), which will be explained in Chapter 3. A lower RMSE value corresponds to a better fit. If we compare the fit of the probability density functions of the three distributions currently used in the GasDay lab, we see that we were getting good estimates of the one-in-N low temperature

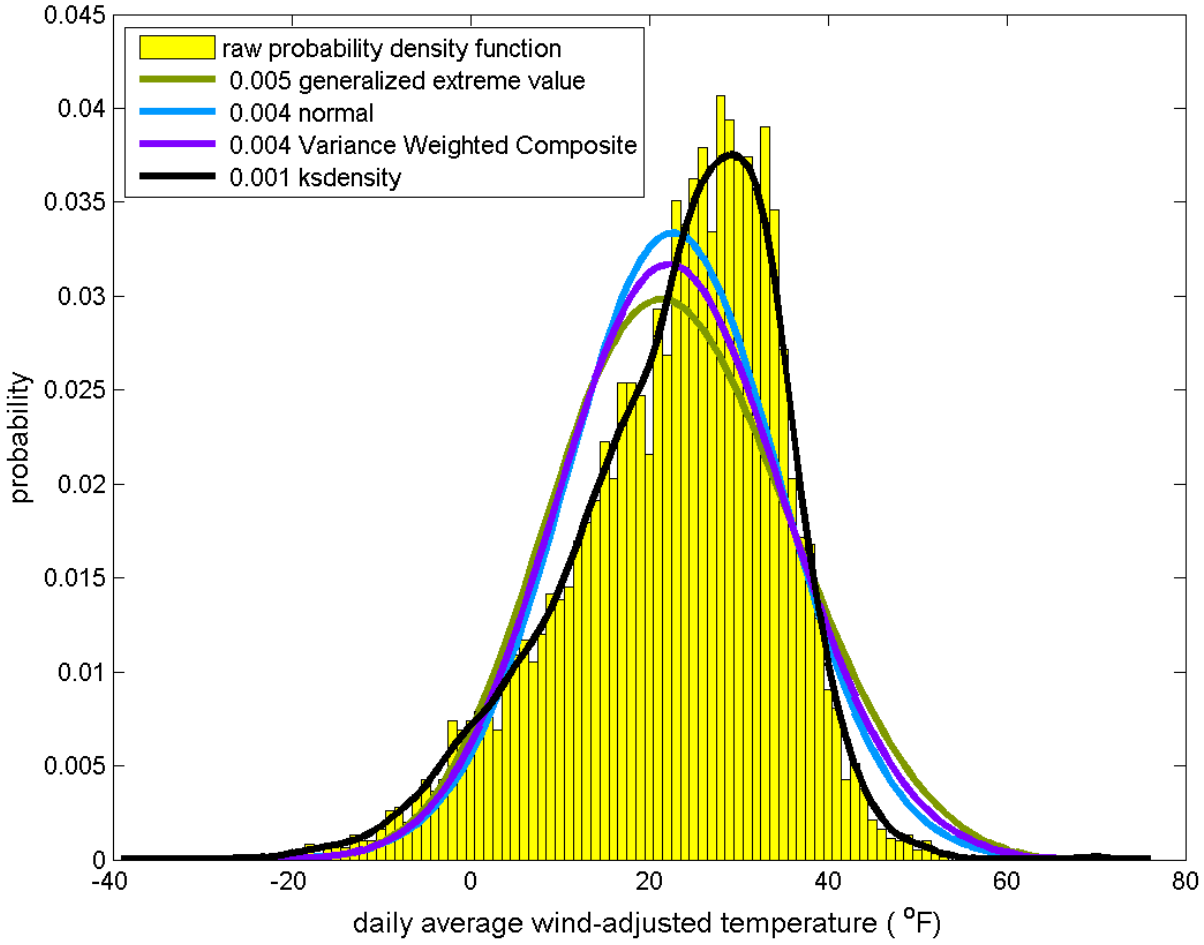


Figure 1.2: Daily temperature for typical winter days for every year in the data set

thresholds. However, there is always room for improvement in science. We found that we could improve these estimates for the one-in-N low temperature threshold by using a non-parametric distribution called the kernel density estimation method. Since the kernel density estimation method makes no prior assumptions that the data comes from a specific distribution, it can fit the weather data well, especially in the left tail. Hence, this thesis will estimate the one-in-N low temperature threshold

for different weather stations in the United States using the non-parametric distribution method, kernel density estimation. Extensive preliminary tests were conducted comparing the results obtained from the different distributions, and the conclusion was reached that the kernel density estimation method provides a better estimate compared to the other currently used distributions.

To summarize the preliminary analysis of Figure 1.2, the figure displays the generalized extreme value distribution, the normal distribution, the variance-weighted composite distribution, and the kernel density estimate. Visually, a good fit would be how closely the probability density function follows the data, determined by a low RMSE value. In Chapter 3, we will explain what we define as typical winter days and the other ways in which we prepared the data for this thesis. We will also re-introduce Figure 1.2 in Chapter 3 so that the reader can better understand this graph.

1.4 Thesis Outline

This chapter provides the reader with a brief overview of the thesis. We explain the motivation of the thesis and provide an informal explanation of the problem we are trying to solve. Then we provide a formal problem statement complete with equations. An introduction to the possible solution naturally follows as the next section, where we provide the reader with the data used and the methods currently

used to solve this problem. Finally, we introduce the nonparametric distribution method called the kernel density estimation method. We glance at the results obtained from the one-in-N algorithm and visually determine that it was an improvement over the current methods. We conclude this chapter by promising the reader more details on the one-in-N algorithm and results.

The remainder of the thesis is organized as follows.

Chapter 2 of this thesis provides a summary of the literature surveyed across disciplines to develop the one-in-N algorithm. This literature survey investigates other methods used by scientists to model the tails of distributions. We explain some theory of the statistical distributions used in this thesis and provide additional references for the reader's benefit. This information prepares the reader for the explanation of the method in Chapter 3.

Chapter 3 describes the mathematical model used to estimate the one-in-N year low temperature threshold. We give a detailed, step-by-step development of the kernel density estimation method algorithm. We first explain how the the data is prepared and explain what a "winter" means in this thesis. We provide the reader with the output of the algorithm and explain how the one-in-N method is developed. We also provide a high-level summary of the entire method to aid in replication of the work.

Chapter 4 provides an evaluation of these results based on extensive tests. We discuss the results obtained from the one-in-N algorithm for Milwaukee, WI (KMKE), Albuquerque, NM (KABQ), and Anchorage, AK (PANC). We describe the test plan used to show that the one-in-N algorithm is an improvement over the other methods, and present the results of the test plan. We hope that these results will encourage additional research in this subject.

This thesis concludes in Chapter 5 with a summary of method and results, as well as suggestions for future research work that originated from this thesis. We also identify some bias in the kernel density estimation method and how it is mitigated in this thesis. The future work is subdivided into two categories: Extension of work and Hypotheses to be explored.

CHAPTER 2

Literature Survey on the Likelihood of Rare Events and Statistical Theory

Extreme events rarely occur and are found in the tails of a statistical distribution. Applications of statistical methods in the estimation of extreme events are seen in fields including meteorology, actuarial sciences, social sciences, economics, business, and engineering. A survey of literature from some of these disciplines reveals a few techniques that may be used to estimate extreme low-temperature events, which are the subject of this thesis.

2.1 Survey of Rare Events in Meteorology, Ecology, and the Nuclear Power Industry

In meteorology, a group of climate scientists, social scientists, and biologists met to discuss the impact of extreme weather and climate events and attempted to discern whether these events were changing in frequency or intensity [2]. Together, they published a series of five articles to discuss these effects. Meehl [2] introduces the concept of extreme events and how a change in the mean and variance of a climate variable affects the frequency of occurrence of these extreme events. Figures 2.1, 2.2,

and 2.3 represent a normal distribution of a climate variable [2]. The shaded areas located in the tails of the distributions depict the extreme events that occur infrequently. Figure 2.1 shows that an increase in the mean affects the frequency of the extreme events; the frequency of the right-tailed extreme events increased; and those of the left-tailed extreme events decreased. Figure 2.2 shows that an increase in the variance increases the frequency of the extreme events. Figure 2.3 shows that a change in both the mean and the variance alters the occurrence of extreme events. Meehl further states that it is possible to estimate changes in extremes that occur once every $10 - 100$ years using extreme value distributions such as the Gumbel distribution. However, in Chapter 3, we will see that the Gumbel distribution, a special case of the generalized extreme value distribution, does not provide a good fit for the left tail of the daily minimum temperatures compared to the kernel density estimation method.

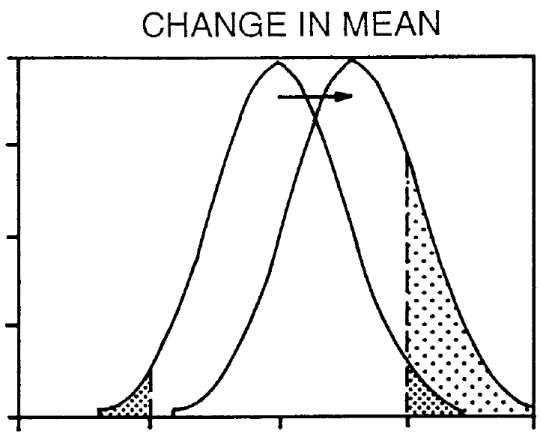


Figure 2.1: Schematic diagram depicting how a change in mean can affect extreme events [2]

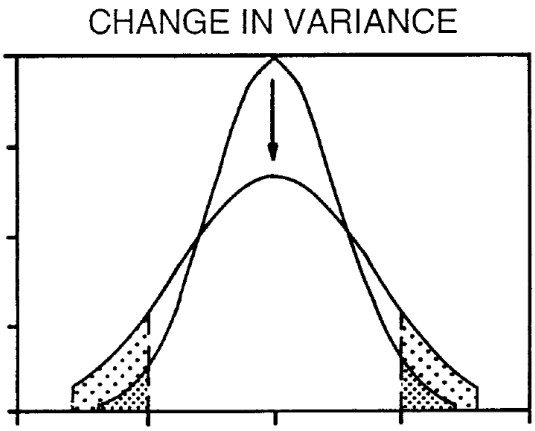


Figure 2.2: Schematic diagram depicting how a change in variance can affect extreme events [2]

Mearns [6] studies the likelihood of extreme high-temperature events and their effect on agriculture in or near the U.S. Corn Belt, including weather stations in the states of Iowa, North Dakota, and Indiana. With her expertise in meteorology, she expects that changes in the means of meteorological variables (e.g.,

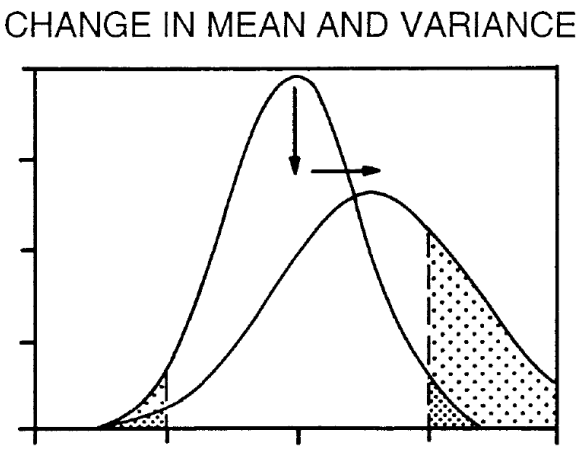


Figure 2.3: Schematic diagram depicting how a change in mean and variance can affect extreme events [2]

temperature) will have adverse effects on agricultural production. She studies the probabilities of the following events:

1. Maximum temperature on a given day in July $\geq$ threshold temperature (day event)
2. At least one run in July consisting of at least five consecutive days $\geq$ threshold temperature (run event)
3. At least five days in July (not necessarily consecutive) $\geq$ threshold temperature (total event)

The threshold temperature is set to 95 °Fahrenheit because it represents the approximate temperature that is reported to be harmful to the corn crop. Her study analyzes extreme high-temperature events for the month of July, as this is the month when a particularly temperature-sensitive agricultural process takes place. The run event (described as event 2 above) is particularly important, as it is argued theoretically to be more harmful to the crop. Her study analyzes how a change in the mean, variance, and autocorrelation of the daily maximum time series data affects the probabilities of the aforementioned events. For this analysis, she has to develop a probabilistic model that simulates the daily maximum time series data. To develop this model, she needs to obtain several characteristics of the time series data such as the shape of the distribution, measure of central tendency (e.g., the mean), measure of dispersion (e.g., the variance), and a measure of the dependence among the data points (e.g., the autocorrelation function). She uses a normal probability density function to obtain the sample mean, sample variance, and sample first-order autocorrelation from the high-temperature time series data for July. The available sample data ranges from 31 to 69 years. She assumes that the daily maximum time series data is an approximate realization from a first-order autoregressive [denoted AR(1)] process or a "Markov" process, which assumes that the data comes from a normal probability density function. She simulates 500 years of July daily maximum time series data using the AR(1) model. Then she varies the

parameters in six different ways to observe how these different changes affect the probabilities of the three extreme events listed above.

The conclusions from her experiments support her hypothesis that a small change in the mean maximum temperature causes shifts of practical significance in the probabilities of extreme high-temperature events. For example, a 3°F increase in the mean, holding the variance and autocorrelation constant, causes the likelihood of occurrence of the run event to be about three times greater than that under the current climate at Des Moines, and the likelihood increases to as much as six times greater when the variance and autocorrelation are increased as well. For the purpose of this study, Mearns' assumption that the daily maximum temperature time series data follows a normal distribution is a good approximation, as it is computationally less intensive to generate synthetic time series data, and it also serves the purpose of trying to analyze the impact that a change in mean maximum temperature has on the likelihood of the three extreme events. However, in this thesis, we are trying to estimate the threshold temperature for the event that the minimum temperature on a winter day is less than or equal to the threshold temperature, which may be expected to occur, on average, one-in-N years. For this purpose, the normal probability density function and hence, the AR(1) model, is not a good approximation for the daily minimum temperature time series data, as it does not model the tails of the data (where the extreme temperatures are located)

very accurately. As a result, much valuable information about the extreme temperatures is lost. The kernel density estimation method is found to provide a better fit for extreme minimum temperatures.

In his paper considering modeling extremes in projections of future climate change, Gerald Meehl [7] summarizes the knowledge of possible future changes in the statistical aspects of weather and climate extremes based on existing models published in a recognized meteorological report. He discusses several climate variables including temperature, precipitation, extratropical storms, El Niño Southern Oscillation (ENSO), and tropical cyclones. We will only discuss the section concerning temperature. Meehl's review of existing climate models leads him to conclude that the weather and climate extremes in a future climate are affected by an increase of greenhouse gases as theoretically expected by meteorologists. For example, an increase in mean temperatures results in higher frequencies of extreme high temperatures and lesser frequencies of extreme low temperatures. Another conclusion is that the diurnal temperature range is reduced due to the observation of a dramatic increase in nighttime low temperatures compared to the daytime high temperatures in many regions. A third observation about the change in temperature extremes is from a decreased daily variability of temperature in winter and an increased variability in summer in the Northern hemisphere. However, the final conclusion is not necessarily one that we have observed with actual winter daily

average temperatures. In Albuquerque, NM, we have observed that there is **more** volatility in the frequency of occurrence of the minimum extreme temperatures. This observation will be discussed further in Chapter 4. Nonetheless, in general, Meehl's conclusions provide an interesting point of view that suggests possible areas for future research based on work reported in this thesis.

Alexander Gershunov [8] analyzes the influence of El Niño Southern Oscillations (ENSO) on intraseasonal extreme rainfall and temperature frequencies in the United States. We will only discuss the results obtained for temperature. Gershunov uses a compositing technique that he developed [8] to demonstrate ENSO sensitivity in the extreme ranges of a temperature probability density function. He conducted his experiment on 168 weather stations in the contiguous United States using six decades of daily data. He found that ENSO-based predictability is potentially useful to predict extreme warm temperature frequency in the southern and eastern United States during El Niño winters and in the Midwest during the strongest events. Extreme warm temperature frequency is very well predicted by La Niña winters in southern United States centered on Texas. However, extreme cold temperature frequency predictability is mostly weak and inconsistent, particularly during strong ENSO events. However, during weaker El Niño winters, this predictability improves in the northern United States, along the West Coast, and in the Southeast. Weaker La Niña winters improve extreme cold temperature

frequency predictability in the Midwest. This paper also suggests that El Niño and La Niña regional specifications are not opposites of each other. The conclusions from Gershunov's work is useful for future work on extreme cold temperature trends in the United States and how ENSO sensitivity affects the winters.

The paper by Qiqi Lu et al. [9] informs our analysis using the kernel density estimation method. Lu discusses general trends in weather over a few centuries of data. It was written as an improvement to an existing method. Lu and her team found a simple but effective way to handle changepoints of weather stations, when there is a change of station location, station instrument, or station shelter. Observations between changepoints are termed as a "regime" [10]. She uses a simple linear regression model to fit the weather data and uses an ordinary least squares method to estimate the trend parameters. She then uses a nonparametric local averaging smoother in conjunction with geographic information system software to plot the trends on contour maps. Lu analyzes weather by month to estimate the weather trends. She encountered missing observations within the weather data from all the stations. The missing weather data was infilled using a model-based expectation maximization algorithm [9]. An interesting conclusion from this research is that the variability of the estimated trends is the greatest during winter and smallest during fall and summer. This paper also shows that the winters in the U.S. show the most warming compared to the other seasons. Specifically, there is

warming in the northern Midwest and the Four Corners region, the Dakotas, southern Arizona, and southern California. In this thesis, we look for similar trends in the weather.

Paul Knappenberger [11] discusses the daily temperature trends in the United States during the 20^{th} century. He discovers three different periods of change: warming from $1900 - 1940$, cooling from $1940 - 1969$, and warming from $1970 - 1997$. From his analysis of the temperature data, he finds higher extreme maxima in the first period, lower extreme minima in the second period, and warming of the extreme minima in the third period. He concludes that the warming of the coldest days of the year in this last period (a period of the greatest human alterations on the climate) is evidence of temperature moderation. He also points out that the high temperatures in this period remain comparatively unchanged. For this study, he uses daily temperature data because most extreme events occur on a fine temporal scale, so using monthly data may overlook many important aspects of how the change took place. In this thesis, we use daily temperature for the aforementioned reason and because we have acquired good quality daily temperature data in the GasDay lab from sources such as the National Oceanic and Atmospheric Administration, NOAA (`http://weather.noaa.gov/weather/WI_cc_us.html`), Schneider Electric (`http://www.schneider-electric.com/`), and the Agricultural Weather Information Service, AWIS

(http://www.awis.com/Forecast_services/About_Forecast_Services.htm). One could use the United States Historical Climatology Network (HCN) [12] as a source of data, although data quality and control is not as strict as it is from the other sites. A constant challenge in obtaining clean data is that the reduction of data quality and control is associated with time of observation changes, station changepoints (location/instrument), and urbanization. Another source of data quality reduction to note is that in the 1980's, liquid-in-glass thermometers were replaced by thermistor-based temperature observing systems. If Knappenberger's data was missing fewer than ten observations, he interpolated the missing values as the linear average between the temperature on the previous and following days. If more than ten observations were missing, he dropped that year of data. In this thesis, we do not want to reduce further the quality of data, so we drop the missing observations. He also makes a point about urbanization in his paper. Urbanization increases night-time temperatures more than day-time temperatures, leading to apparent increasing trends in minimum temperatures. D. R. Easterling [13] also looks at extreme climate trends worldwide. He observes that in some areas of the world, increases in extreme events are apparent, but in others, there seems to be a decline. This paper makes important contributions regarding the trends in temperature in the U.S., but the data set only extends until 1997. On February 2^{nd}, 2011, both Milwaukee and Albuquerque experienced an extreme cold event, which might be evidence that the climate is not tending towards moderation anymore, as

suggested by Knappenberger. Further analysis of this event can be found in Chapter 4 of this thesis. Easterling [13] suggests that an increased ability to monitor and detect multidecadal variations and trends is critical to detect changes in trends and to understand their origins.

The research conducted by Rebetez [14] for two weather stations in Switzerland led to insightful findings regarding temperature variability in Europe. He found that warmer temperatures are attributed to a decrease in day-to-day temperature variability (measured using intra-monthly standard deviation of temperature), particularly for minimum temperatures and winter. He also found that a negative correlation exists between the day-to-day variability and skewness of the temperature distribution. This means that a reduction in the day-to-day variability occurs through the loss of the coldest extremes in the monthly distribution, particularly the coldest extremes in winter. He also attributes a warming climate to the reduction in diurnal temperature range, i.e., a reduced warming of daytime temperature compared to nighttime temperature. This observation is particularly prevalent at lower elevations. Rebetez discusses the effect of a meteorological phenomenon called the North Atlantic Oscillation (NAO) index, which is the dominant mode of winter climate variability in northern Asia and in the North Atlantic region spanning North America and Europe. The NAO is an atmospheric mass that seesaws between the subtropical high and polar low. This

index varies over the years, but may remain in one phase for several years at a time. A positive NAO index indicates a subtropical high pressure center and a strong Icelandic low, leading to cold and dry winters in northern Canada and Greenland and mild and wet winters in the eastern United States [15]. Rebetez observed that higher NAO index values are associated with an increase in temperature and a decrease in day-to-day temperature variability. This is consistent with the fact that these high pressures are linked to high NAO values and relatively stable weather in winter. In this thesis, we are concerned with estimating the cold event that falls in the coldest extremes of the winter distribution. In future research, we will be interested to see if these extreme cold events change over time. It also may be interesting to evaluate the effect of the NAO index and elevation on the extreme cold temperatures. However, in this research, we observe that the temperature variability decreases in the summer and is higher in winter. Hence, it is important to define the window of days to be narrow enough to avoid the low variability and broad enough to have a significant number of data points as we discuss in Chapter 3.

A study of the trends in time-varying percentiles of daily minimum temperature over North America reveals a unique warming pattern of the daily minimum temperature [16]. In this study, Robeson analyzes percentiles ranging from the 5^{th} to the 95^{th} in $5-$percentile increments that were estimated for each month of every year using linear interpolation. Then linear trends were estimated

using a least-squares regression followed by interpolation using splines. Cluster analysis was used to identify regions with homogeneous percentile trends over the year. An average-linkage method was used to identify larger homogeneous clusters. High quality data was used spanning the years $1948 - 2000$ with less than 20% of missing data.

He found three principal spatial patterns for the daily minimum temperature, with two of the three patterns that were dominant (covered 95% of North America). One cluster is found in eastern North America and shows moderate warming trends during February and March, but very weak trends during the other months. Another cluster is found in western North America and shows intense warming during January through April. However, the lower tail of the daily minimum temperature frequency distribution had the strongest warming for the lower percentiles from January through March. He also found that during the other parts of the year, trends in daily minimum temperature are mostly positive, with weak cooling occurring during October and November. The last cluster, found in a small part of northeastern Canada, has strong to moderate cooling during the colder months and weak warming in warmer months. These trends in daily air temperature percentiles emphasize the importance of late winter and spring in the changing climate of North America. However, the data set used in this study only spanned until the year 2000. The year 2011 had some unusually cold weather on

February 2^{nd} (discussed in Chapter 4), which may impact the results of this study. From our analysis in this thesis, Albuquerque, NM, seems to fall within the first cluster, the one with moderate warming, which is in accordance with our findings from the results charts in Chapter 4. Further research into the results reported in this thesis may be useful in recognizing trends similar to those in Robeson's work.

In Great Britain, F. K. Lyness [17] was interested in being able to meet the gas demands for a very cold winter that may occur with a frequency of once in fifty years. He breaks down the problem into two parts:

1. what constitutes 1-in-50 winter conditions?
2. what is the demand for gas in these conditions?

We will only discuss the first part of the problem here. Lyness uses a span of at least 51 winters to deduce the 1-in-50 winter conditions. He realizes that making estimates from a sample size of 51 involves a large sampling error, but a longer historical span of data either may not be available, would raise the problem of climatic trends, or both. He performed tests of randomness and found that although climatic changes can be detected, they follow no predictable pattern and are very small compared to the seemingly random variation from winter to winter. Another problem is that meteorologists are unclear about how to include climatic effects in the process of making estimates. Since Lyness intends to update his estimates every

$5-10$ winters, including climatic effects may be unnecessary. He also realizes that the results he obtained for the 1-in-50 winter estimates are sensitive to particularly cold or mild winters and the size of the sample. His justification is that we cannot know the "true" 1-in-50 value and that he is attempting to solve a technical problem practically. His aim was to find a practical and consistent approach for all the natural gas regions in Great Britain. Our aim is to find a distribution that best models the left (cold) tails of the daily winter temperature distribution, while being practical to use. The method Lyness uses is unique. He first chooses a series of temperature threshold values at random. Then the accumulated temperature below each threshold is calculated for each of the 51 winters. He then fits a probability distribution to the 51 accumulated temperatures and finds the 1 in 50 value from the distribution. He found that a normal distribution fitted to the cube-root of accumulated temperature fits the data well. However, the existence of zeros in the data causes problems, so they were dropped from the sample. In this thesis, we can provide a possible improvement to these estimates as we do not need to eliminate zero values of temperature. Also, the kernel density nonparametric distribution might be a better solution compared to the cube-root normal distribution as it does not make any parametric assumptions to the data.

Sebastian Jaimungal [18] investigates the use of kernel-based copula processes (KCPs) to analyze multiple time-series and to model interdependency

across multiple time-series. A copula function is a joint distribution function of uniform random variables. He applies the theory to daily maximum temperature series from weather stations across the United States. He successfully modeled the heteroskedasticity of the individual temperature changes and discovered interdependencies among different weather stations. He illustrated the superior modeling power of KCPs by comparing the models obtained from KCPs with those from a Gaussian copula process. He points out that KCPs handle missing data naturally. In his application, he detrended the temperature data by subtracting a customized sinusoidal seasonal trend, based on a least-square criterion, from the data. We use a Fourier series process to remove seasonality of the temperature data and focus on its stochastic nature. He then analyzes the second moment autocorrelation function (autocorrelation of the square of the data). This second moment autocorrelation function implies the rate of fluctuation, or volatility of the temperature data. Analyzing the second and the first moment autocorrelation functions, he developed a non-stationary kernel function. From this paper, we are further convinced that using kernel density function is a good solution to modeling the cold tail of the temperature probability density function.

From the ecological society, Philip Dixon [19] attempts to improve the precision of estimates of the frequency of rare events. The probability of a rare event is estimated as the number of times the event occurs divided by the total sample

size. However, this estimate has very low precision, and the coefficient of variation (cv) of this estimate can exceed 300% for sample sizes smaller than 100 observations. The coefficient of variation is the normalized measure of dispersion of a probability distribution. If σ is the sample standard deviation, and μ is the sample mean,

$$cv = \frac{\sigma}{\mid \mu \mid}; \qquad \mu \neq 0 \; . \tag{2.1}$$

To reduce the cv to below 10%, one should obtain sample sizes of $10^3 - 10^4$ observations. He explains that since such a large number of observations are not always available, auxiliary data should be used to improve the precision of the estimate. He describes four approaches for creating auxiliary data: (1) Bayesian analysis that includes prior information about the probability; (2) Stratification; (3) regression models; (4) using aggregated data collected at larger spatial or temporal scales. He applied these methods to data on the probability of capture of vespulid wasps by the insectivorous plant *Darlingtonia californica*. He found that all four methods increased the precision of the estimate compared to the simple frequency-based estimate. In this thesis, we do not use auxiliary data in our estimation process. However, future research in auxiliary data for temperature can be analyzed to determine if it improves the estimates. Then we may be able to increase the amount of data we have to about a century.

Another application of statistical methods is in system reliability. Reliability

theory is used to estimate a safety circuit failure of a nuclear reactor by Babik [20]. A reactor safety circuit is a complex device whose function is to trip a nuclear reactor when it develops a dangerous condition. Shutting down a faulty reactor is the purpose of the safety circuit, but it can fail to perform this task. Babik derives formulae for the frequency of occurrence of safe and unsafe circuit failures using reliability theory. However, reliability theory is used to determine the probability of encountering a failure [21]. The binary nature of whether a system fails does not exactly meet the requirements for estimating the extreme low-temperature events, since the latter is not binary. This problem can be overcome by determining a threshold value below which is considered an extreme low-temperature event. However, this solution does not exactly solve the problem of finding the extreme low-temperature event with a probability of one-in-N years, but in this thesis we will use a similar binary method to test our one-in-N algorithm.

The United States Department of Agriculture created a "Plant Hardiness Zone Map" which displays the average annual minimum temperature in the United States, Mexico, and Canada [22]. This average annual minimum temperature is based on at least 10 years of temperature data. Cathey used weather data from 8000 weather stations to create this map. As an extension of the work presented in this thesis, we could use a similar mapping technique to display the extreme

low-temperature threshold in the United States to aid in the visual analysis of the extreme cold temperature trends.

In conclusion, a survey of literature across several disciplines has uncovered different strategies to help estimate the extreme low-temperature event. In the next section, we will discuss statistical theory for the distributions used in this thesis.

2.2 Parametric and Nonparametric Statistical Theory

In this section, we will discuss the statistical theory for the distributions used in this thesis. These statistical distributions include the Gaussian (normal), the log normal, the generalized extreme value, and the kernel density estimation method. We will also provide a brief definition of probability density functions and cumulative distribution functions.

2.2.1 Probability Density Functions

A function with values *f(X)*, defined over the set of all real numbers, is called a probability density function (pdf) [23] of the continuous random variable X if

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx \qquad \text{for any real constants } a \text{ and } b. \tag{2.2}$$

A probability density function, integrated from a to b (with $a \leq b$), gives the probability that the corresponding random variable will have a value in the interval from a to b. Also, the value of the probability density function of X at a is zero in the case of continuous random variables. The total area ($-\infty < x < \infty$) under the probability density function curve is equal to 1. Figure 2.4 is an example of a probability density function. Next, we will define cumulative distribution functions and provide an example graph.

2.2.2 Cumulative Distribution Functions

If X is a continuous random variable, and the value of its probability density at t is *f(t)*, then the cumulative distribution function (cdf) [23] of X is

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\, dt \qquad \text{for } -\infty < x < \infty\,. \tag{2.3}$$

One important property of a cumulative distribution function is $F(\infty) = 1$. An example of a cumulative distribution function is shown in Figure 2.5.

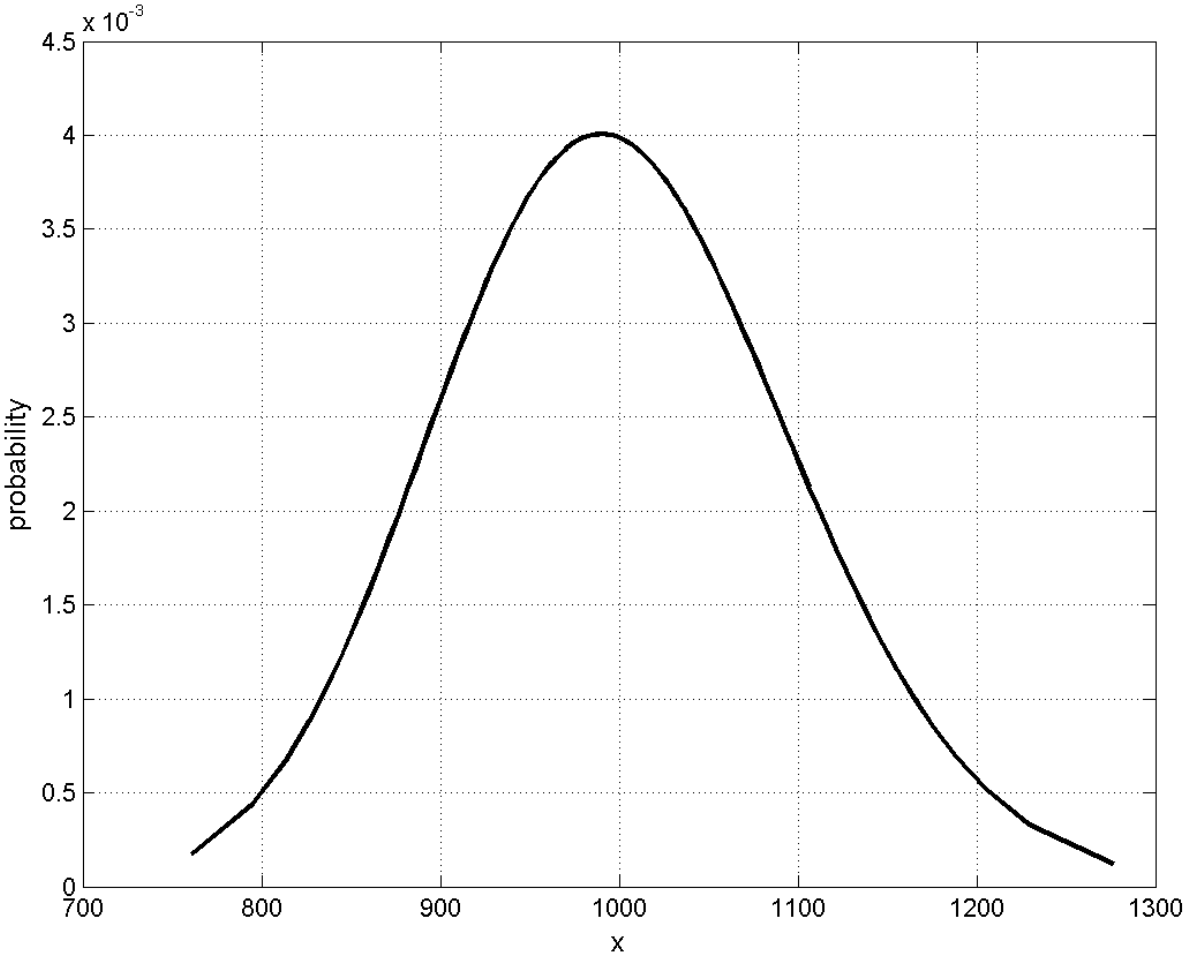


Figure 2.4: Example of a probability density function

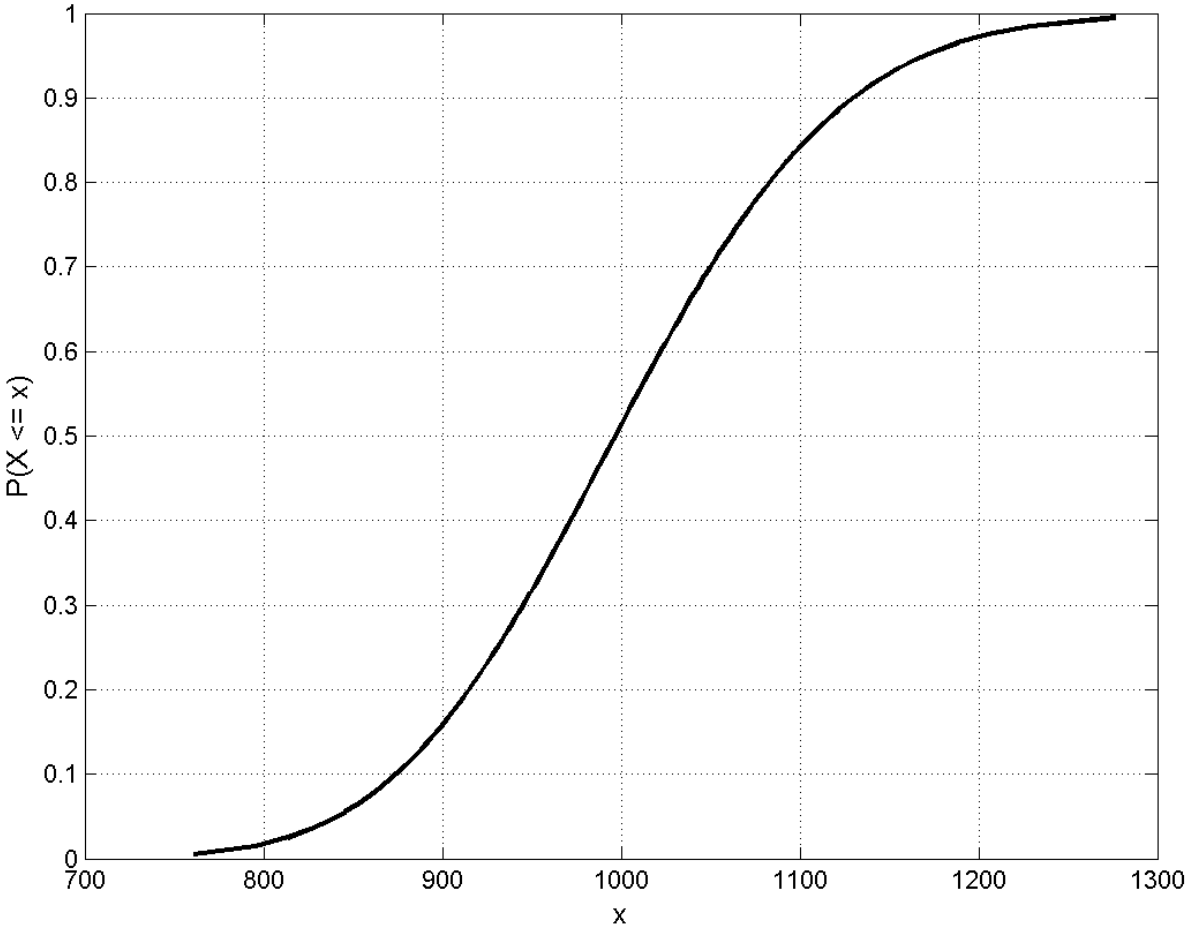


Figure 2.5: Example of a cumulative distribution function

2.2.3 Gaussian (Normal) Distribution

A random variable X with mean μ and standard deviation σ has a normal distribution if its probability density function is

$$f(x \,|\, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \qquad \text{for } -\infty < x < \infty, \quad \text{where } \sigma > 0. \tag{2.4}$$

The normal distribution was first studied by Abraham de Moivre, Pierre Laplace, and Karl Gauss [24]. From Figure 2.6, we see that the probability density function of the normal distribution looks like a cross section of a bell and is sometimes referred to as a bell curve. Figure 2.7 depicts the cumulative density function of a normal distribution function. We will perform a comparative analysis of the threshold temperature obtained using the normal cumulative distribution function with the one obtained using the kernel density estimation method in Chapter 4.

2.2.4 Log Normal Distribution

The normal and log normal distributions are closely related [23]. A random variable X with mean m and standard deviation s has a log normal distribution as shown in Figure 2.8 if its probability density function is

$$f(x \,|\, m, s) = \frac{1}{sx\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{\ln(x)-m}{s}\right)^2}, \qquad \text{for } 0 < x < \infty\,. \tag{2.5}$$

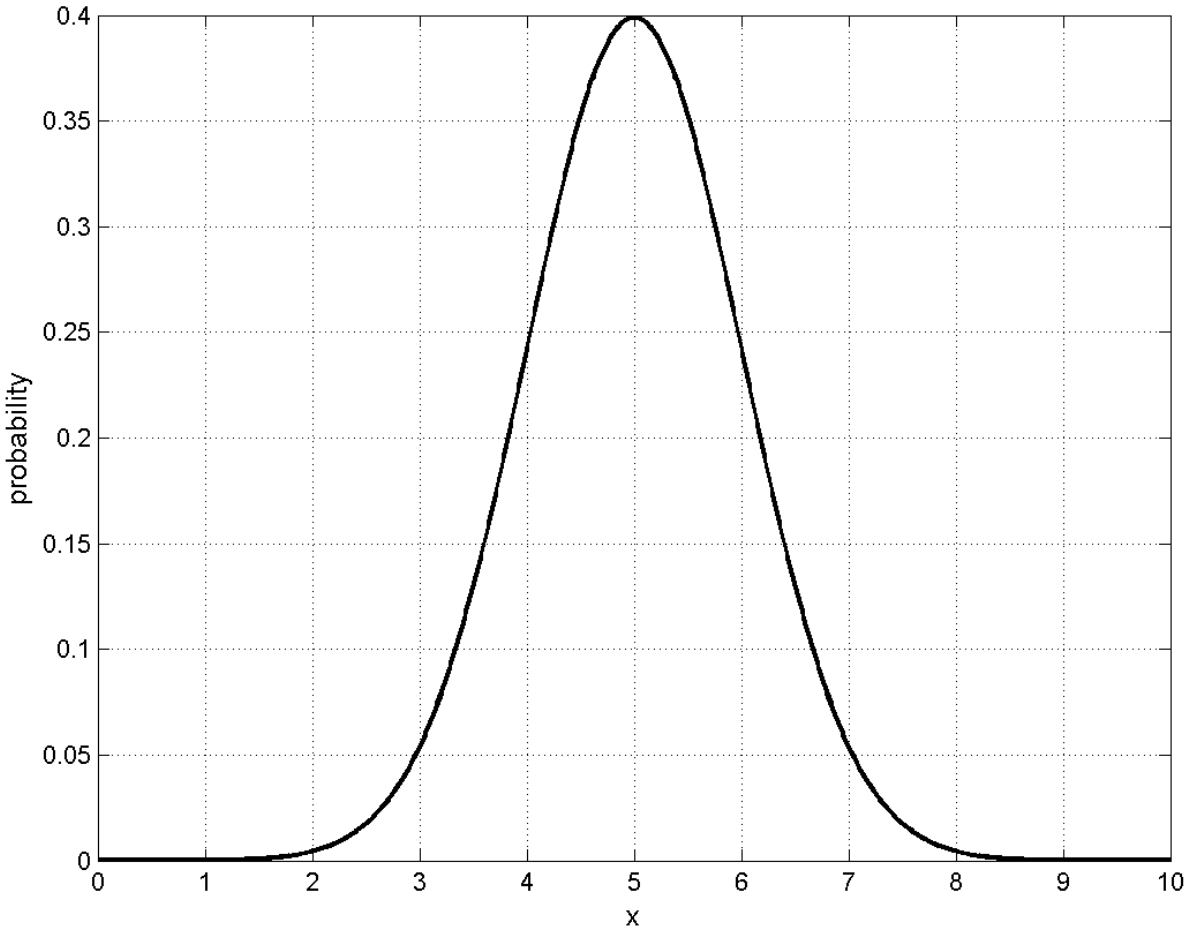


Figure 2.6: Gaussian (normal) probability density function

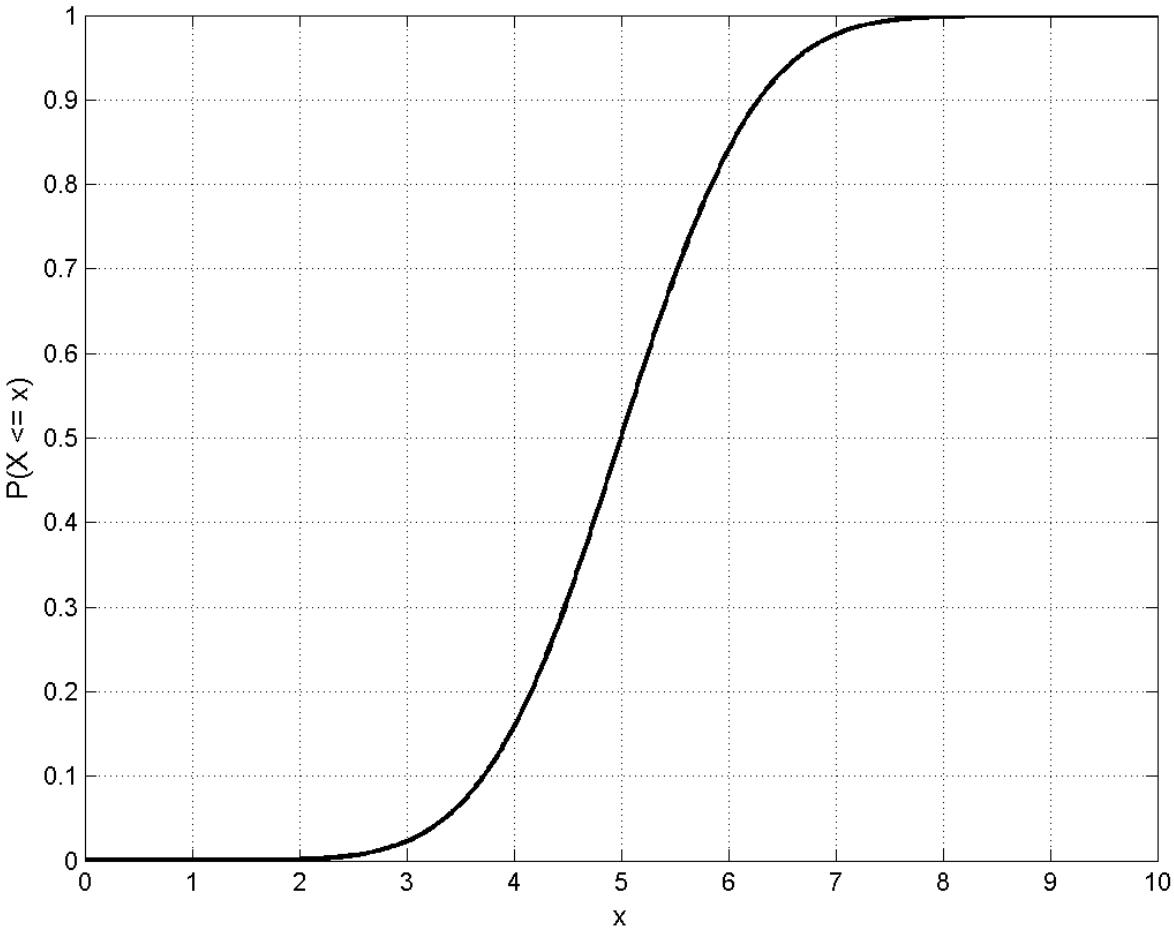


Figure 2.7: Gaussian (normal) cumulative distribution function

Then $\ln(x)$ is distributed normally with mean μ and standard deviation σ as shown,

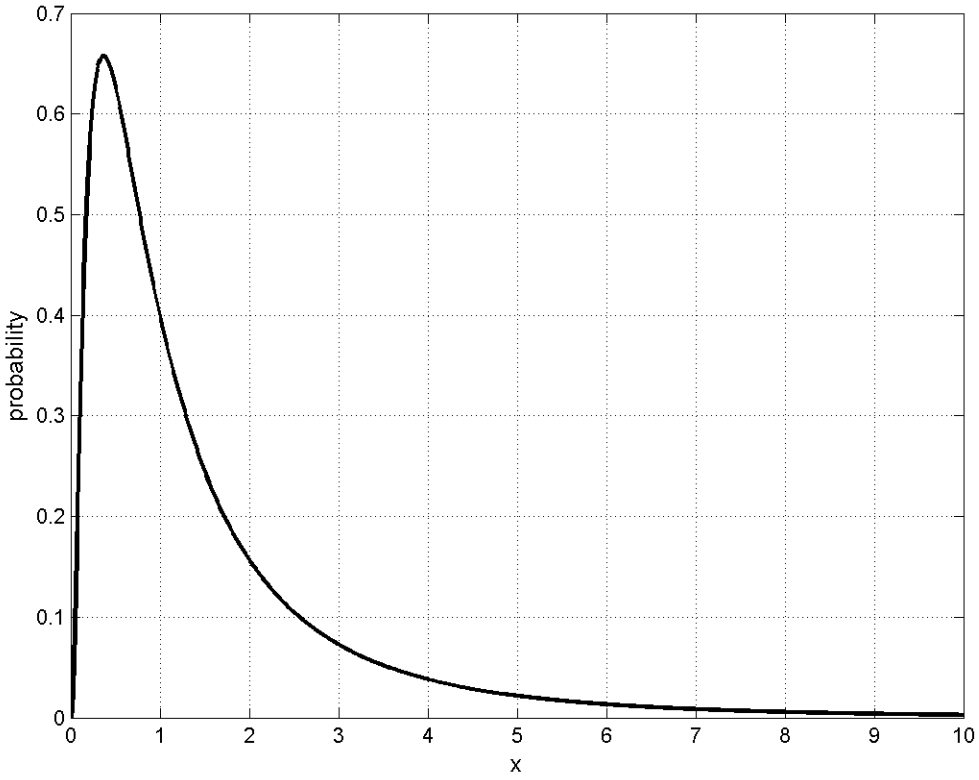


Figure 2.8: Log normal probability density function

$$\mu = e^{\frac{(m+s^2)}{2}} \; ;$$

$$\sigma^2 = e^{(2m+s^2)}(e^{s^2} - 1) \; . \tag{2.6}$$

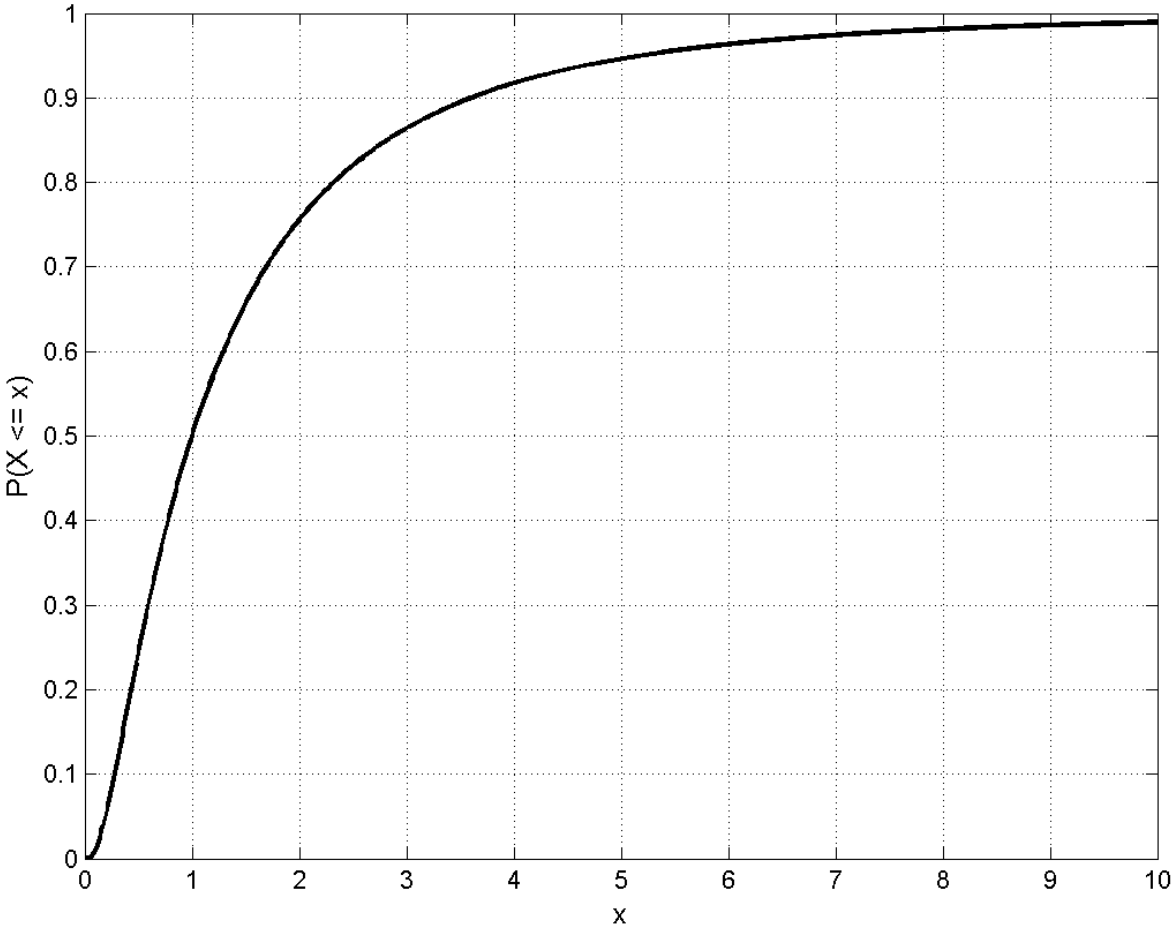


Figure 2.9: Log normal cumulative distribution function

From Figure 2.8, we see that the left tail is finite and non-negative. One LDC uses a log normal distribution in an attempt to fit the cold tail of the winter temperature data. A graph of the cumulative distribution function of the log normal distribution is displayed in Figure 2.9.

2.2.5 Generalized Extreme Value Distribution

A random variable X with shape parameter ξ, location parameter μ, and scale parameter σ has a generalized extreme value distribution if its probability density function is [25]

$$f(x|\xi,\mu,\sigma) = \begin{cases} \frac{1}{\sigma}\left\{1+\xi(\frac{x-\mu}{\sigma})\right\}^{-\frac{1}{\xi}-1} e^{-(1+\xi\frac{(x-\mu)}{\sigma})^{-\frac{1}{\xi}}} & \text{if } -\infty < \text{x} \leq \mu - \frac{\sigma}{\xi} \text{ for } \xi < 0; \\ & \text{and if } \mu - \frac{\sigma}{\xi} \leq \text{x} < \infty \text{ for } \xi > 0; \\ \frac{1}{\sigma} e^{(-\frac{(x-\mu)}{\sigma} - e^{-\frac{(x-\mu)}{\sigma}})} & \text{if } -\infty \leq \text{x} < \infty \text{ for } \xi > 0. \end{cases} \tag{2.7}$$

The generalized extreme value distribution was first introduced by Jenkinson [25]. The shape parameter ξ (K in Figures 2.10 and 2.11) may be used to model a wide range of tail behavior. The case $\xi = 0$ (Type I) depicts an exponentially decreasing tail in the probability density function. An example of this distribution is the Gumbel distribution as shown in Figure 2.10 as Type I. The case $\xi > 0$ (Type II) corresponds to a long-tail in the probability density function. An example of this

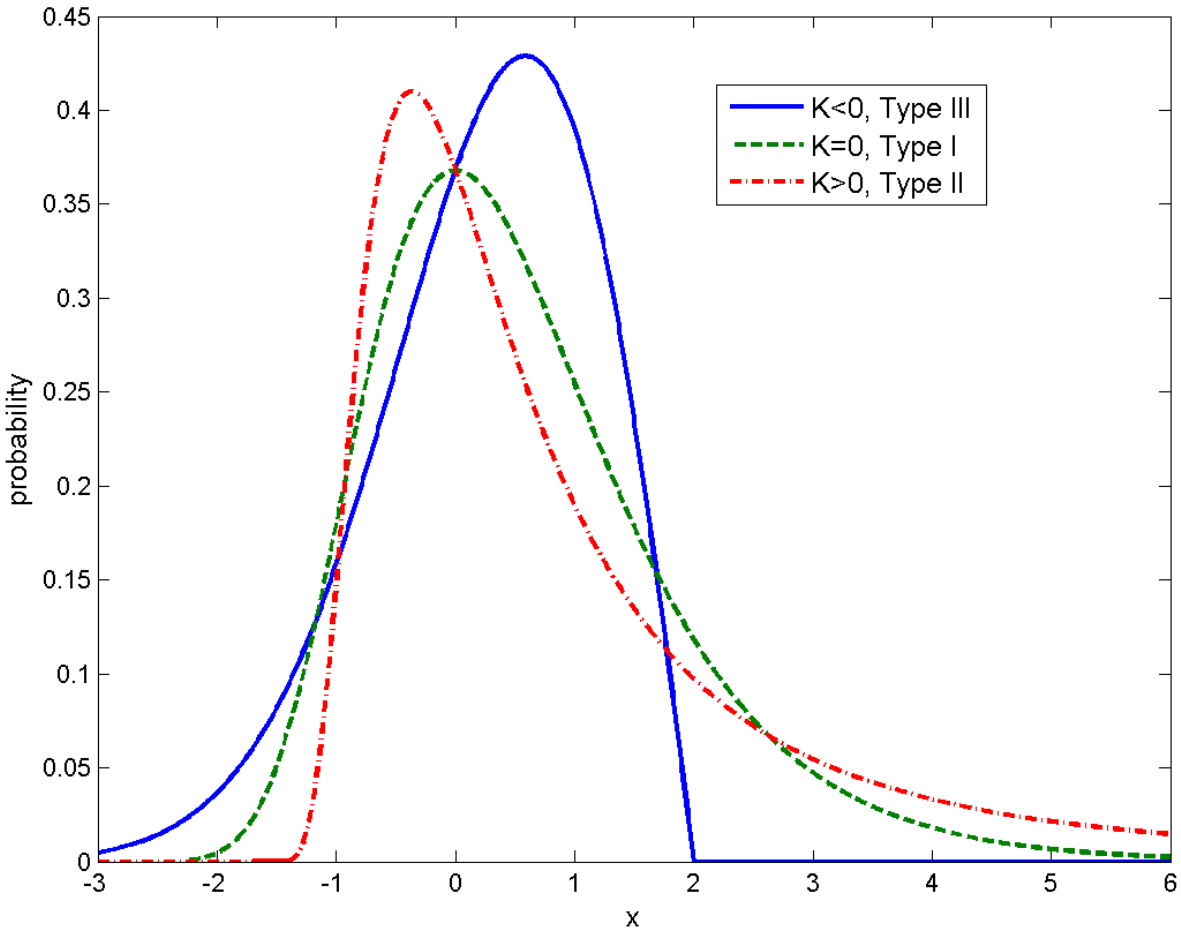


Figure 2.10: Three types of the generalized extreme value probability density function

distribution is the Fréchet distribution displayed as Type II in Figure 2.10. The case $\xi < 0$ (Type III) depicts a short tail in the probability density function because it has a finite upper endpoint. An example of this distribution is the Weibull distribution shown in Figure 2.10 as Type III. The associated cumulative distribution functions for all three types of the generalized extreme value distribution can be found in Figure 2.11. The Weibull distribution has a longer left

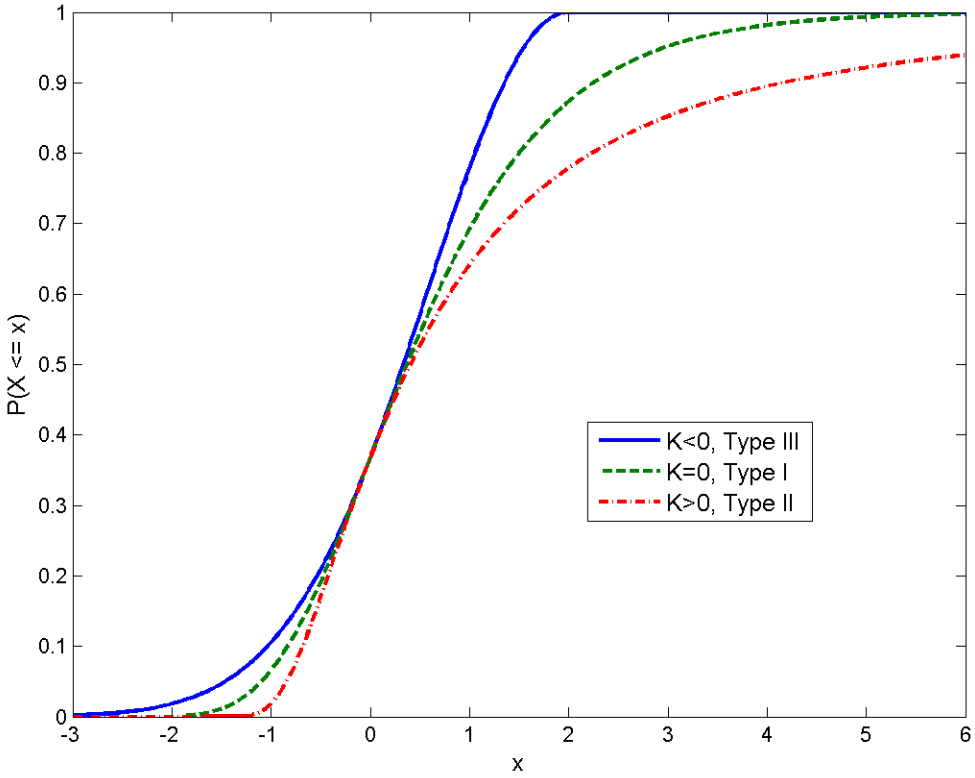


Figure 2.11: Three types of the generalized extreme value cumulative distribution function

tail, which has been used to model the cold tail of the winter temperature data to obtain the one-in-N threshold temperature (Chapters 3 and 4). We will discuss the analysis of the generalized extreme value distribution further in Chapter 4. Applications of the generalized extreme value distribution can be found in Paul Embrechts' book on *Modelling Extremal Events* [26]. In the next section, we will explain some theory of the kernel density estimation method.

2.2.6 Kernel Density Estimation

A probability density function for independent and identically distributed random variables X can be estimated using the kernel density estimation method [3]

$$\widehat{f_h}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) , \tag{2.8}$$

where $K(\cdot)$ is a kernel function, $h > 0$ is a smoothing parameter called the bandwidth, and n is the sample size. Table 2.1 displays a few of the commonly used Kernels. I(...) corresponds to the indicator function or a characteristic function defined on a set X that indicates membership of an element in a subset A of X. For example, the uniform kernel assigns a weight of 1 for each observation that falls into the interval $[x - h, x + h)$ and a weight of 0 for all observations outside this interval. In this thesis, we use the Gaussian kernel function.

The bandwidth (window parameter) h controls the smoothness of the probability density function estimate. Hence, it is crucial to choose an appropriate bandwidth. On one hand, if the bandwidth is too small, the result is a crude estimate of the probability density function. On the other hand, of the bandwidth is too large, then we get an overly smoothed estimate of the probability density function. In this thesis, we will use Silverman's rule of thumb [3] to estimate a

Table 2.1: Kernel Functions

Kernel	K(u)
Uniform	$\frac{1}{2}I(\|u\| \leq 1)$
Epanechnikov	$\frac{3}{4}(1-u^2)I(\|u\| \leq 1)$
Quartic (Biweight)	$\frac{15}{16}(1-u^2)^2I(\|u\| \leq 1)$
Gaussian (Normal)	$\frac{1}{\sqrt{2\pi}}e^{(-\frac{1}{2}u^2)}$

practical bandwidth. This window parameter is optimal for a normal distribution. If σ is the calculated standard deviation, and n is the total number of points in the data set, then the bandwidth is

$$h = \sigma \left(\frac{4}{3n}\right)^{\frac{1}{5}} . \tag{2.9}$$

From Figure 2.12, we can gain insight as to how the kernel density estimation method works [3]. There is a kernel function centered at each of the observations. At a given x, we find the probability density function estimate by vertically summing over the kernel "bumps" [3]. This explanation also helps understand how varying the bandwidth changes the appearance of the bumps and the appearance of their sum.

Figures 2.13 and 2.14 display the probability density function and cumulative

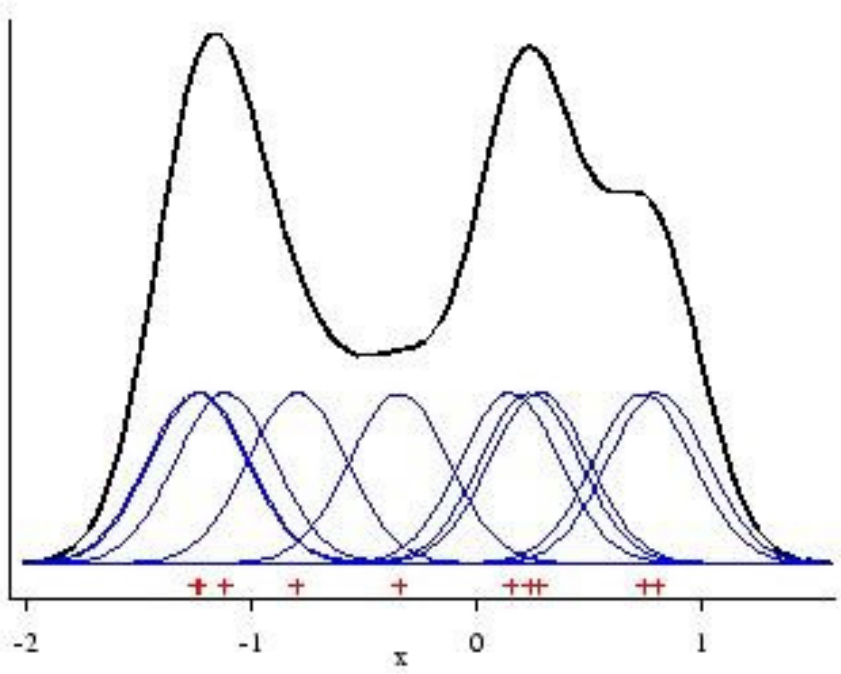


Figure 2.12: Cartoon of Construction of the Kernel Density Function [3]

distribution function, respectively, for data obtained from a gamma distribution and how closely the kernel estimation matches the original gamma distribution.

2.2.7 Variance-weighted composite distribution

In this thesis, we introduce a variance-weighted composite distribution that is being used to fit the cold tail of the winter temperature data. The variance-weighted composite distribution was created using a weight determined from the variance of different distributions. In practice, this composite is created using the normal,

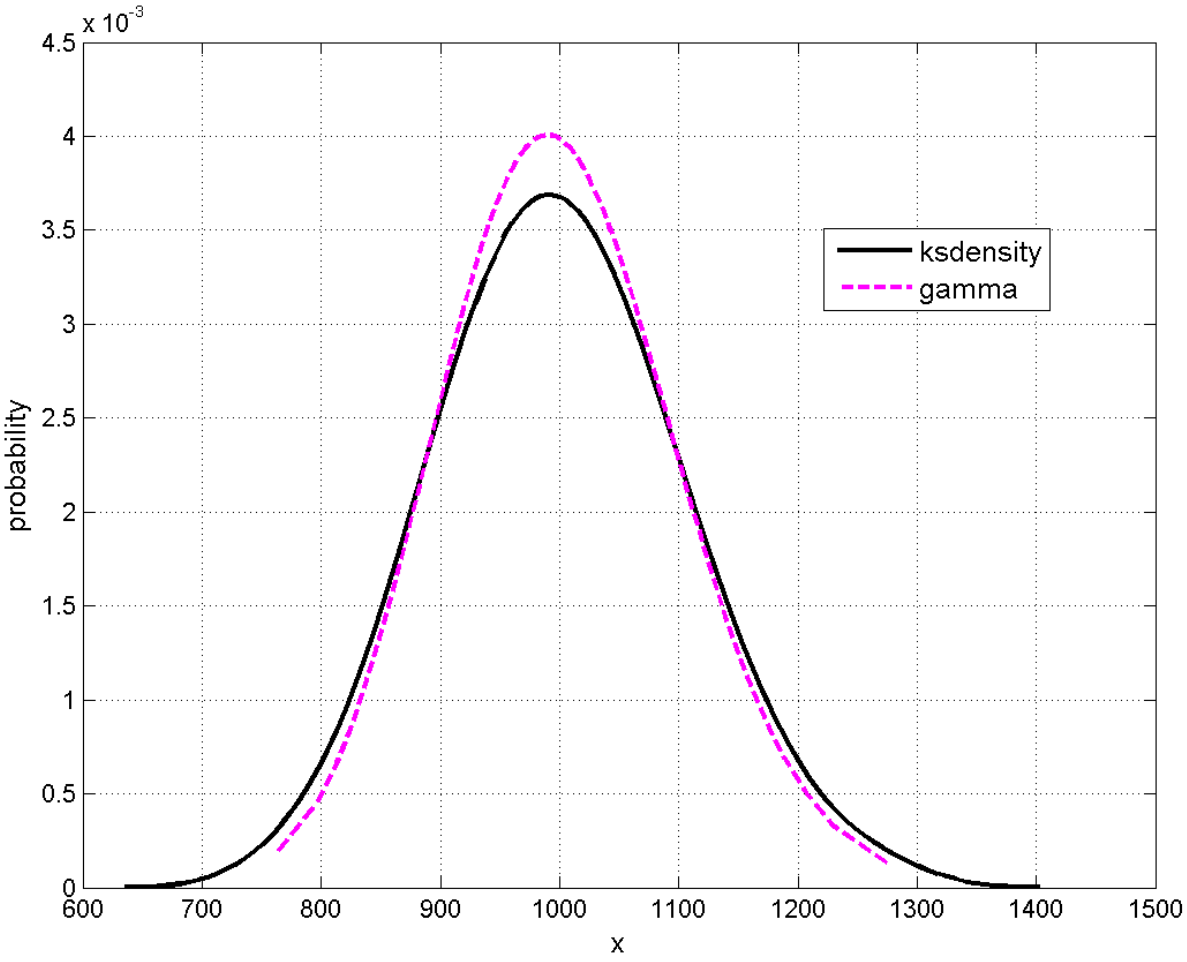


Figure 2.13: Comparison between the Gamma probability density function and its Kernel Estimate

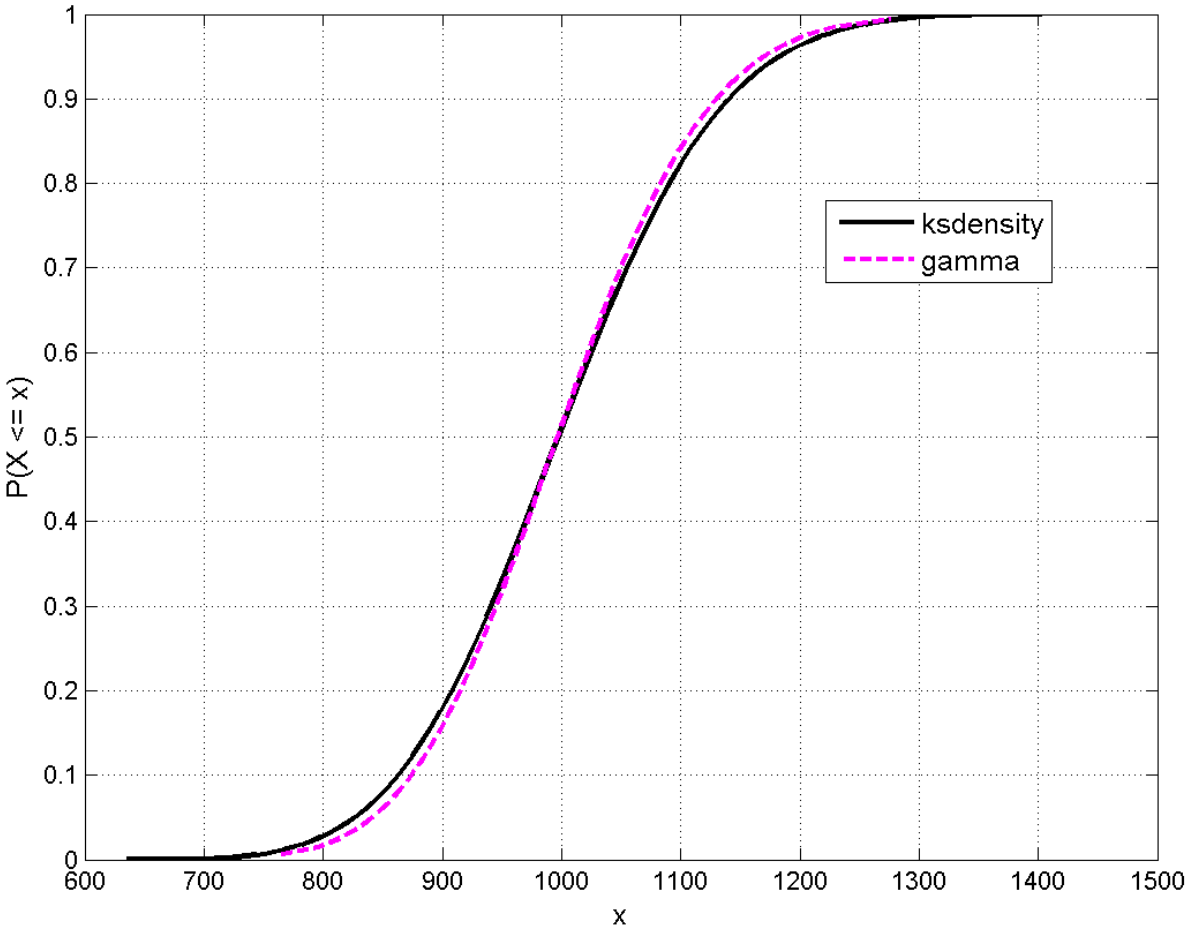


Figure 2.14: Comparison between the Gamma cumulative distribution function and its Kernel Estimate

Weibull, Gumbel, and generalized extreme value distributions (see Section 1.3, Chapter 1). For the purpose of this thesis, we will only discuss a composite created using the normal and generalized extreme value distributions because in Chapter 4, we will test the one-in-N algorithm and compare the results obtained with the test results obtained using this variance-weighted composite.

The variance-weighted composite distribution is created by first summing the reciprocal of the variances of the two distributions. We define this sum as `WeightDivisor` as shown in Equation 2.10.

$$\texttt{WeightDivisor} = \frac{1}{\text{variance}_{\text{normal}}} + \frac{1}{\text{variance}_{\text{GEV}}} \,. \tag{2.10}$$

The reciprocal of the variances of each of the two distributions are divided by the `WeightDivisor` yielding in a 1×2 matrix

$$\texttt{weight} = \frac{1}{\texttt{WeightDivisor}} * \left[\frac{1}{\text{variance}_{\text{normal}}} \qquad \frac{1}{\text{variance}_{\text{GEV}}} \right] \,. \tag{2.11}$$

The resulting variance-weighted composite distribution is the product of the `weight` and the probability density function matrix or the cumulative distribution function matrix of the normal and GEV distributions, in this example.

In this chapter, we have presented a review of techniques used to evaluate extreme events in various applications, particularly in the field of meteorology. We

also have explained statistical theory of several distributions that are being used currently to estimate the low temperature threshold, as well as the kernel density estimation method, which is the statistical method proposed to estimate the one-in-N threshold temperature. In Chapter 3, we present the method used to estimate the one-in-N coldest threshold temperature for the winter. This method is an improvement over the existing methods used.

CHAPTER 3

Estimating the One-in-N Coldest Temperature Threshold

In Chapter 2, we presented a review of several techniques used in evaluating extreme events, particularly in the field of meteorology. We also provided a brief discussion of statistical theory for the different distributions that will be used in this thesis. Armed with this background knowledge, now we are prepared to explain what data we used, how it was obtained, and how it was prepared for use in the one-in-N algorithm. Part of the data preparation stage is defining what a "winter" means in this thesis. We also will show the resulting output of the one-in-N algorithm, so the reader has a better understanding of the one-in-N algorithm used. Following the output section, we will explain the one-in-N algorithm. We will close this chapter with a brief discussion summarizing the main points of this chapter. Table 3.1 defines most notation used in this chapter.

3.1 Summary of Problem

In this section, we review the problem statement and the contribution made by this thesis in the field of meteorology and statistics.

Table 3.1: Common notation used in Chapter 3

Notation	Description
X_t	daily average wind-adjusted temperature (°F)
n	total number of years in the data set
t	index of days {1, 2, 3, ..., $n * 365$}
T_{th}	one-in-N low temperature threshold
N	0.25, 0.5, 1, 2, 5, 10, 20, and 30 years

In this thesis, we estimate the threshold defining an extreme cold temperature event that may be expected to occur, on average, once in N years for different weather stations in the United States using the non-parametric statistical method called kernel density estimation.

From Chapter 1 Section 1.2, we reiterate the assumption of stationarity of weather and climate. Since we intend to update our estimates every year, including climatic effects might be unnecessary [17]. We also assume that daily temperature is independent of neighboring days, but not identically distributed. The remainder of this section is a reproduction of Chapter 1, Section 1.2, for the reader's convenience.

We let X_t be a random variable describing the daily average wind-adjusted temperature on a day, where the domain of the random variable is the set of all days in the entire historical record, and the range is the set of all possible temperatures. For our experiment, we use n years, then t is the index of days {1, 2,

3, ..., $n \times 365\}$. We estimate a threshold temperature, T_{th}, with the property that the event $X_t \leq T_{th}$, may be expected to occur, on average, once in N years. We call this threshold T_{th} the "one-in-N low temperature threshold." We consider $N =$ 0.25, 0.5, 1, 2, 5, 10, 20, and 30. Below, we discuss the properties of the one-in-N low temperature threshold.

We define an indicator function, $f_c(X_t)$, whose domain is temperature (X_t), and whose range is the set $\{0, 1\}$. The rule is

$$f_c(X_t) = \begin{cases} 1 & \text{if } X_t \leq T_{th} \ , \\ 0 & \text{otherwise} \ . \end{cases} \tag{3.1}$$

Our experiment is to count the number of times the temperature (X_t) falls below the threshold temperature (T_{th}). Our outcome is the independent random variable, *count* (Equation 3.2), which is the count of the number of events in an n-year period,

$$count = \sum_{t=1}^{n \times 365} f_c(X_t) \ . \tag{3.2}$$

Hence, if we perform several experiments with different sets of n years of data, the expected value of *count* should be

$$E\left(count\right) = E\left(\sum_{t=1}^{n \times 365} f_c(X_t)\right) = \frac{n}{N} \ . \tag{3.3}$$

For example, if we had $n = 300$ years and $N = 30$ years, then if we perform several experiments of counting the number of times the temperature falls below the one-in-N low temperature threshold, each with 300 years of data, we should get $E(count) = 10$.

A survey of the literature in Chapter 2 reveals that there is an opportunity to improve the low temperature threshold estimate. In the GasDay lab, we use parametric distributions to obtain low temperature threshold estimates (Chapter 2). In this thesis, we use the non-parametric distribution called the kernel density estimation method to obtain the low temperature threshold estimate, T_{th}, that occurs on average, once in N years. This method is not only simple to use but also models the cold tails of the data better than the distributions currently used. The cartoon in Figure 3.1 gives a high-level summary of steps in the construction of the one-in-N algorithm.

3.2 Preparing the Data

In this section, we explain how the data are prepared for the one-in-N algorithm. Figure 3.2 shows the daily average wind-adjusted temperature data. The data are high quality historical actual daily temperatures used in Marquette University GasDay's forecasting models. These data are first obtained from weather vendors including Schneider Electric (`http://www.schneider-electric.com/`) and the

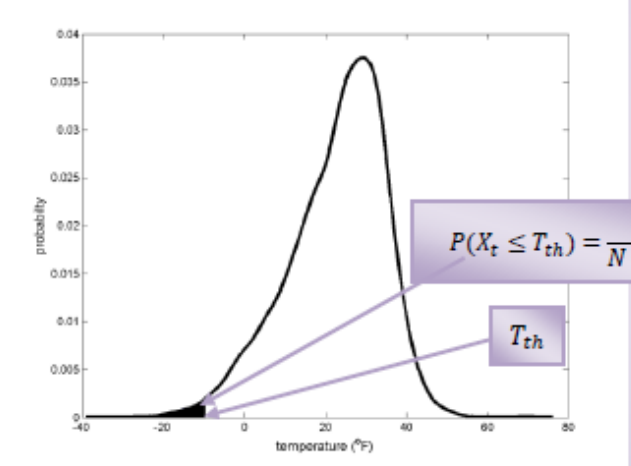
- ***Clean the data***
  - Missing data: Patch missing daily average temperature from reliable weather sources, and discard the days with missing observations that could not be patched
  - For leap years, discard the data for June 30th
  - Wind Speed: Adjust daily average temperature data for wind speed

- ***Calculate the mean***
  - Calculate the mean for the daily average wind-adjusted temperature data
  - Smooth mean daily average wind-adjusted temperature data using Fourier series method

- ***Choose "winter" window***
  - Sort smoothed mean daily average wind-adjusted temperature in ascending order
  - Choose the first 91 mean daily average wind-adjusted temperature data. The 91 days that correspond to this data are defined as "winter" in this thesis.

Estimating the one-in-N year low temperature threshold

- Estimate the probability density function of the cleaned "winter" data using the kernel density estimation method.
- We know that $P(X_t \leq T_{th}) = \frac{1}{N*91}$.
- Using the probability density estimate and the probability of the event $X_t \leq T_{th}$; the one-in-N year low temperature threshold T_{th} is as shown in the figure below



- $P(X_t \leq T_{th})$ is the area under the curve to the left of T_{th}, shaded in black.

Figure 3.1: Steps for the one-in-N algorithm

Agricultural Weather Information Service, AWIS (http://www.awis.com/Forecast_services/About_Forecast_Services.htm). Like all sources of data, the data set obtained from these vendors are missing some observations. We patch the missing data from the National Oceanic and Atmospheric Administration, NOAA (http://weather.noaa.gov/weather/WI_cc_us.html), another reliable source. Using these sources of data, we are able to get sufficiently long (at least longer than 30 years) data sets for many weather stations. However, we still do not have enough data (often less than 10 years) for some weather stations. Another problem is that in spite of having several reliable sources of data, the weather data still contain missing observations. This lack of data is common in applications.

Next, we will explain the process used to clean the data.

Once we have the weather data we need, we identify the weather stations for which we have more than 30 years of daily average temperature data. We need more than 30 years of data because the highest value of N used in this thesis is 30. In this thesis, we use three weather stations, Milwaukee, WI (KMKE), Albuquerque, NM (KABQ), and Anchorage, AK (PANC) for our analysis. All three weather stations have more than 30 years of temperature data (Table 3.2).

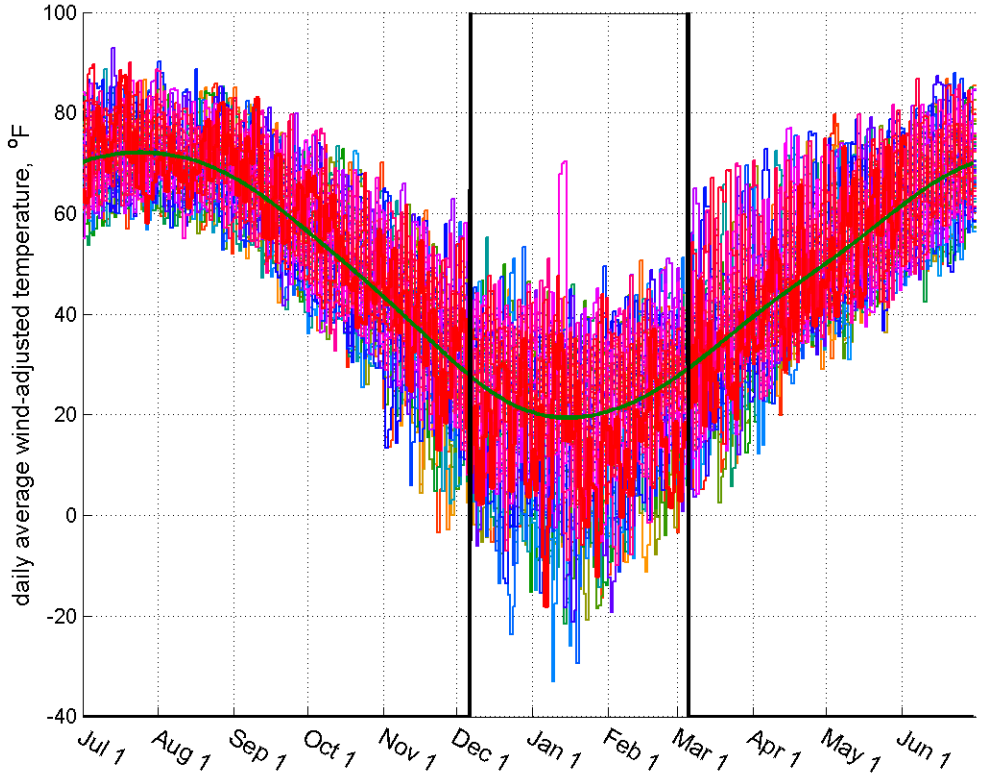


Figure 3.2: Daily average wind-adjusted temperature for Milwaukee, WI (KMKE)

Table 3.2: Size of data set

Weather station	Number of years in data set	Year range
KMKE	67 years	1948 to 2014
KABQ	67 years	1948 to 2014
PANC	42 years	1973 to 2014

Let h be the index for hour, and let temp_h be the temperature at hour h.

The daily average temperature data (DailyAvgTemp) is calculated by averaging all

24 hours of a day's temperatures,

$$\text{DailyAvgTemp} = \frac{\sum_{h=1}^{24} \text{temp}_h}{24} \quad . \tag{3.4}$$

The daily average temperature data is arranged from July 1^{st} through June 30^{th} for each year of data as shown in Figure 3.2. Continuing with the data preparation, n years of data yield an n-by-365 matrix. On leap years, we have an extra day in February, so we drop the June 30th data point to keep consistent with having 365 days in a year. This is the most logical date to omit for this work, as it falls in the summer, and our work is concerned with winter temperatures. Also, from Figure 3.2, we see an interesting event in the middle of January in the year $2002 - 2003$ where there is an unusually high temperature for Milwaukee, WI. This confirms findings of Rebetez [14] and Lu [9] that there is high variability of temperature in the winter.

Continuing with data preparation, we adjust the daily average temperature to account for the effect of wind speed. For this study, wind speed (ws, in miles per hour) is the quantity that affects the rate at which buildings lose heat. A building loses more heat on a windy day compared to a non-windy day at the same temperature [27]. We calculate the Heating Degree Day (HDD65) using a reference

temperature of 65 °F,

$$\text{HDD65} = \max\left(0,\ 65 - \text{DailyAvgTemp}\right) \quad . \tag{3.5}$$

HDD65 is an effective measure for weather in the natural gas industry for the estimation of natural gas consumption [28]. Since heating spaces in homes and businesses is most likely to occur at temperatures below 65 °F, we use a reference temperature of 65 °F. The calculation of HDD65 and reference temperature of 65 °F are natural gas industry standards in the United States [28]. For daily average temperatures less than or equal to 65 °F, we calculate the wind factor (wf) [27]

$$wf = \begin{cases} \frac{152+ws}{160} & \text{if } ws \leq 8; \\ \\ \frac{72+ws}{80} & \text{if } ws > 8. \end{cases} \tag{3.6}$$

For daily average temperatures greater than 65 °F, we do not adjust for wind, so, $wf = 1$. Then the wind-adjusted Heating Degree Day (HDDW65) is

$$\text{HDDW65} = \text{HDD65} * wf \quad . \tag{3.7}$$

We are interested in temperature and not heating degree days in this thesis. To adjust the daily average temperature data for the effect of wind, we calculate the

difference between HDD65 and HDDW65 as

$$\delta\text{HDD65} = \text{HDD65} - \text{HDDW65} \ . \tag{3.8}$$

Then, the daily average wind-adjusted temperature is

$$X_t = \text{DailyAvgTemp} + \delta\text{HDD65} \ . \tag{3.9}$$

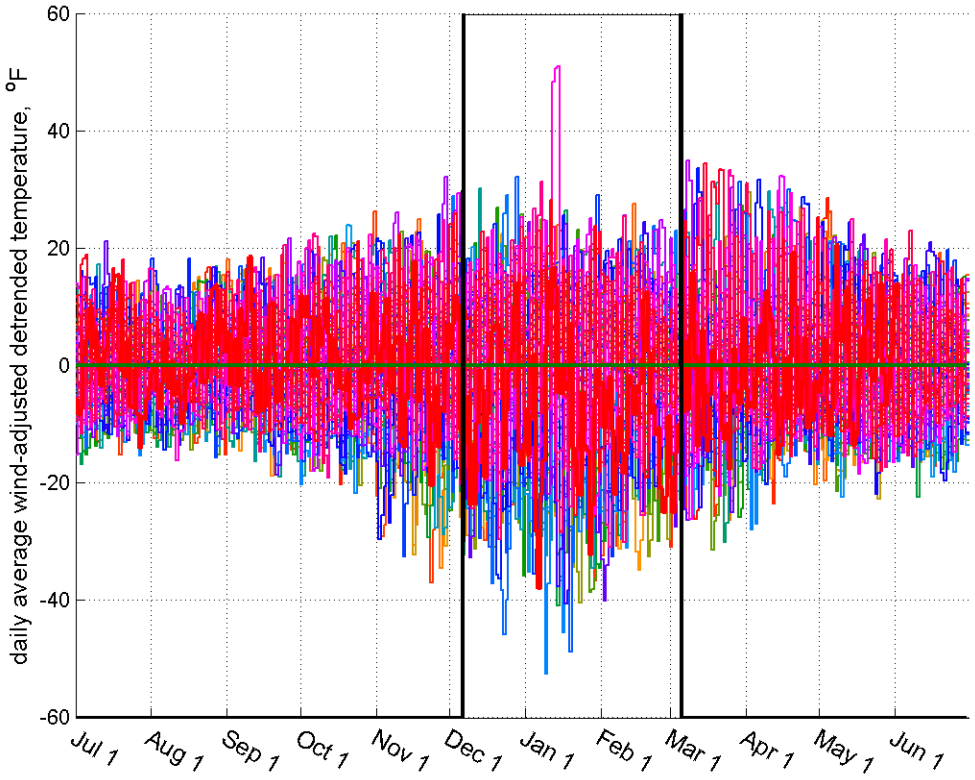


Figure 3.3: Daily average wind-adjusted detrended temperature for Milwaukee, WI

We then find the mean daily average wind-adjusted temperature for each day of the years forming a vector of size one row and 365 columns. The daily average wind-adjusted temperature means are smoothed using a Fourier series, retaining the first five Fourier harmonics. The result is the solid green line in Figure 3.2. From this figure, we can see that the mean is annually periodic. In Chapter 2, we referred to a study done by Jaimungal [18] on the benefits of using kernel-based copula processes. We need to detrend the temperature to focus on the stochastic nature of the data. Hence, we detrend the daily average wind-adjusted temperature data shown in Figure 3.2 by subtracting the smoothed mean from the daily temperature data. The resulting data is defined as detrended (or deviation from normal), and reveals the stochastic nature of the data as shown in Figure 3.3. There is more variability in the data during the winter (located around the center of the figure) compared to the other seasons. What do we define as "winter" in this thesis? We want to choose a window of days that is broad enough to contain a large data set for our analysis, but narrow enough to contain only the data with the most variability, characteristic of winter data.

Choosing the window of days to fit the aforementioned criteria remains a challenge. Empirically, we have seen that a window of 91 days with the coldest daily average wind-adjusted temperatures seems to satisfy this criteria for most weather stations. However, more research into varying the window of days is needed to

improve the estimates of the one-in-N algorithm. We are interested in defining a winter because the problem we are trying to solve in this thesis is estimating the one-in-N coldest threshold temperature that occurs in winter. The window of 91 days of winter is found by sorting the smoothed mean temperature (as described above) in ascending order, so the coldest mean temperature is the first data point in the vector. Then we identify the following 90 coldest mean temperatures and the corresponding days on which all 91 coldest mean temperatures occur. These coldest 91 days are consecutive because of the periodic shape of the temperature data shown in Figure 3.2. These days would not be consecutive if there were a run of unusually high temperatures that are not characteristic of winter weather because then we will see a spike or a more obvious bimodal effect in the smoothed mean. We define bimodal effect as the condition where the second derivative of the mean has more than two inflection points. However, we did not see a bimodal winter effect in any of our weather stations. Hence, the 91 days with the coldest detrended daily average wind-adjusted temperatures are consecutive. The 91 days with the coldest daily average wind-adjusted temperatures are displayed within the box characterized with a thick black line in Figure 3.2, Figure 3.3, and Table 3.3. The 91 days with the coldest daily average wind-adjusted temperatures are wide enough to contain a large data set for our analysis, but narrow enough to only contain the data with the most variability, characteristic of winter data. At this point, the data is cleaned and ready for use in the one-in-N algorithm.

Table 3.3: "Winter" for each weather station assuming non-leap years

Weather station	Start of "winter"	End of "winter"
KMKE	December 6	March 6
KABQ	November 19	February 17
PANC	December 2	March 2

In this section, we have explained how the data was prepared for use in the one-in-N algorithm. We talked about the sources of data and demonstrated how the data was adjusted for wind. We defined a winter as 91 days with the coldest daily average wind-adjusted temperatures, and finally cleaned the data by removing the missing observations. Now, we are ready to use the data in the one-in-N algorithm. Before we discuss the method, we will take a look at the output to get a better idea of how the method works.

3.3 Output of the One-in-N Algorithm

Now that we have prepared the data, the next step is using the data in the algorithm. However, in this section, we will discuss the output obtained from the one-in-N algorithm so that the reader may understand the purpose of the algorithm. We will illustrate the output of the one-in-N algorithm using graphs. We note that the results obtained for the 1-in-N winter estimates are sensitive to

particularly cold or mild winters and the size of the sample. Our justification is that we cannot know the "true" 1-in-N value, and we are attempting to solve a technical problem practically [17].

We estimate the threshold defining an extreme cold temperature event that may be expected to occur, on average, once in N years for different weather stations in the United States using the non-parametric statistical method called kernel density estimation. What is the one-in-N (where $N = 30$) year coldest threshold temperature estimate for the weather station KMKE in Milwaukee? We estimate that it is $-27.1°$F using 67 years of daily average wind-adjusted temperatures in the one-in-N algorithm.

Figure 3.4 displays a histogram of winter data defined by 91 days with the coldest daily average wind-adjusted temperatures for the weather station KMKE. Figure 3.5 illustrates a plot of the raw probability density function created by scaling the histogram by the total number of days ($91 \times n$) in the data set and its estimates from the generalized extreme value distribution, the normal distribution, the variance-weighted composite distribution, and the kernel density estimate. It also shows the RMSE scores in the legend for each distribution. The kernel density estimation method has the lowest RMSE of 0.001, thus providing the best fit to the data. The difference in the mean square errors between the kernel density estimation

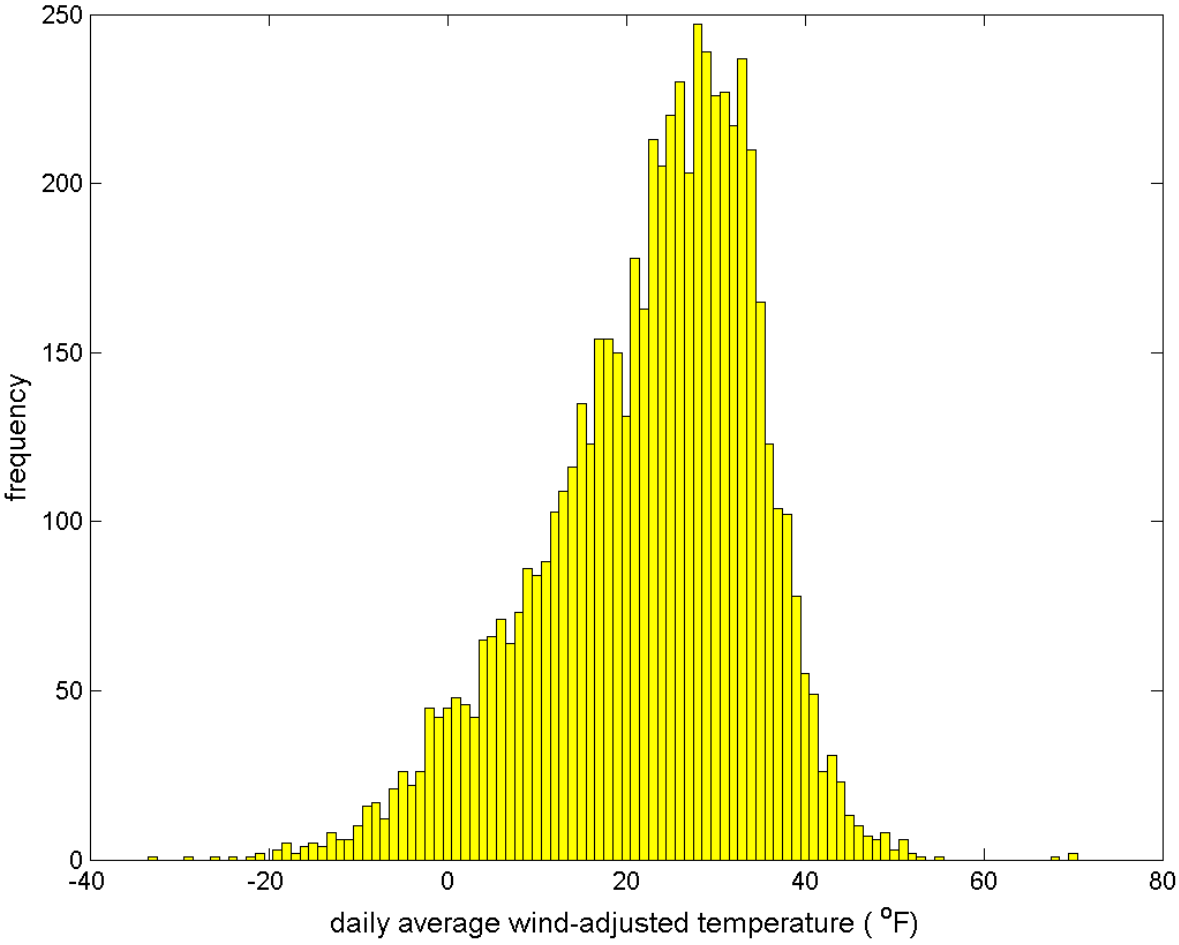


Figure 3.4: KMKE coldest 91 days $\times$ n years temperature histogram

method and each of the other distributions is statistically significant at the 5% level. In the following section, we will explain how the RMSE value is calculated.

3.4 The One-in-N Algorithm

In this section, we explain the details of the one-in-N algorithm. From the data preparation section (Section 3.2), we learned how to adjust the data for use in the

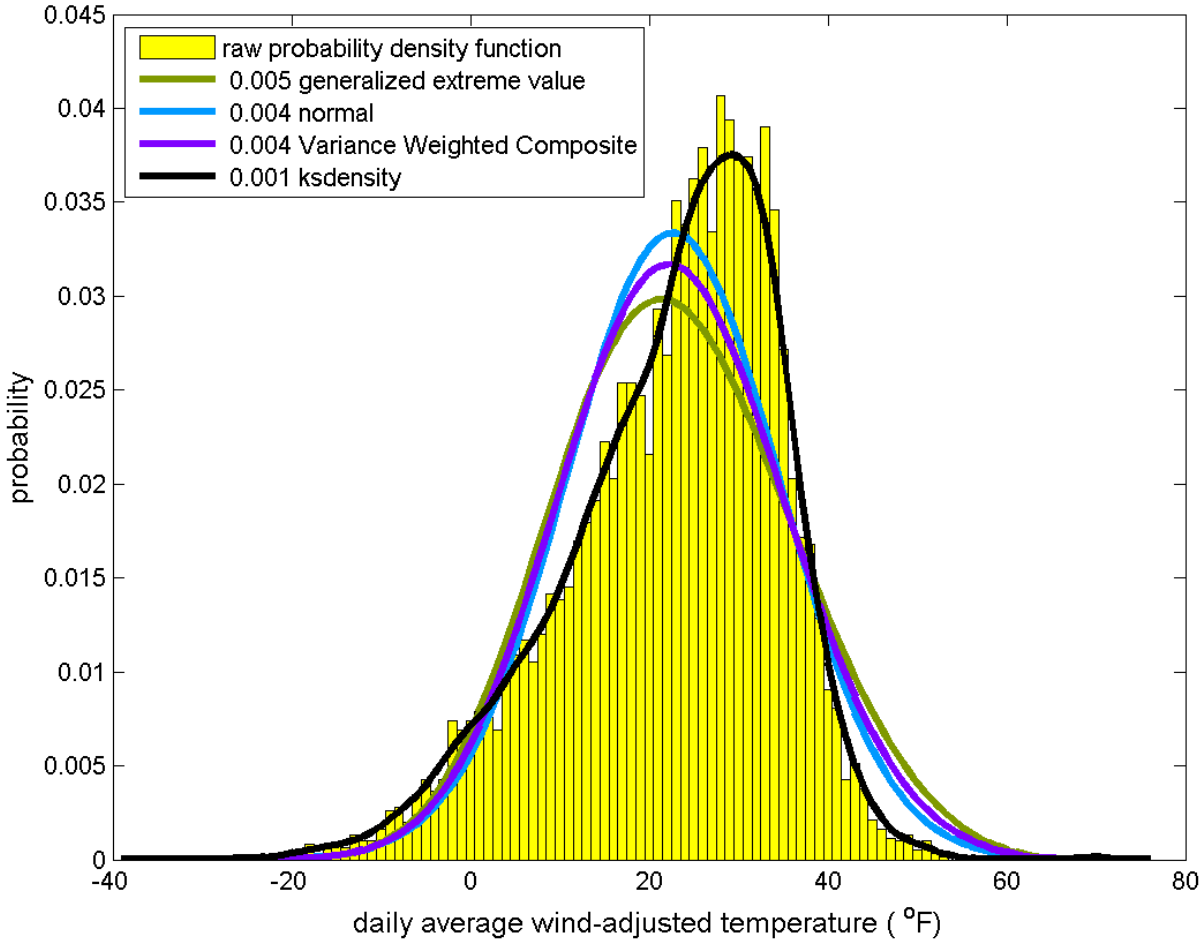


Figure 3.5: KMKE coldest 91 days $\times$ n years temperature with distributions

one-in-N algorithm. In its final form, we have 91 days of coldest daily average wind-adjusted temperature data which we defined as "winter." In Section 3.3, we provided a preview of the output of the one-in-N algorithm. Figure 3.1 is a cartoon explaining the steps of the one-in-N algorithm at a high level.

First, we plot a histogram of the prepared data (for example, Figure 3.4). The histogram gives us a general idea of the spread of the temperature data along

with the frequency of occurrence of temperature during the n years of winter (where n is the total number of years in the data set, and "winter" is 91 days with the coldest daily average wind-adjusted temperatures). Hence, we have $n \times 91$ total data points for the one-in-N algorithm. We set the bin width of the histogram to 1°F for ease of implementation and interpretation. Besides, a bin width of 1°F is broad enough to contain significant number of occurrences of temperature and narrow enough to give the viewer an idea of the general spread of the data. We also store this temperature frequency determined from the histogram for later use (in a vector called `histvalues`). If there is no missing data, we calculate a probability density function scale factor (`pdfscalefactor`) as a product of the total number of years in the data set (n), 91 days with the coldest daily average wind-adjusted temperatures, and the bin width of the histogram (`binwidth`),

$$\texttt{pdfscalefactor} = n * 91 * \texttt{binwidth} \, . \tag{3.10}$$

Second, we estimate the probability density function for the temperature data using the generalized extreme value distribution, the normal distribution, the variance-weighted composite distribution, and the kernel density estimate. We store their respective probability density function values in a vector called `pdfvalues`. We need to measure how well each of these distributions model the data. To do this, we use the `pdfscalefactor` to scale the histogram to a raw probability density

function. Now, the area under the raw probability density function is equal to one. We use `distfitness` as a score to evaluate the goodness of fit of each distribution. This score is essentially the root mean square error (RMSE), calculated as

$$\texttt{distfitness} = \sqrt{\sum_{i=1}^{\text{length(e)}} \frac{(\texttt{e})^2}{\text{length(e)}}} \quad ,$$

$$\text{where} \quad \text{e} = \texttt{pdfvalues}_i - \texttt{histvalues}_i \times \frac{1}{\texttt{pdfscalefactor}} \ . \tag{3.11}$$

From Equation 3.11, it follows that a lower score corresponds to a better fit to the data. Figure 3.5 shows the `distfitness` (RMSE) scores in the legend for each distribution. The value associated with the kernel density estimation method is the lowest and statistically significant at the 5% level. In this figure, we also see the variance-weighted composite distribution created from the normal and GEV distributions (explained in Chapter 2, Section 2.2, Subsection 2.2.7).

Theory suggests that combining distributions sometimes provides a better estimate of the probability density function and the cumulative distribution function [10]. However, in this case, the ksdensity method still seems to provide a better estimate than the variance weighted composite because it has a lower `distfitness` (RMSE) score than that of the variance-weighted composite (Figure 3.5).

Third, we want to calculate the one-in-N coldest temperature threshold.

From Equation 2.2 in Chapter 2, Section 2.2, Subsection 2.2.1, we know that a probability density function, integrated from a to b (with $a \leq b$), gives the probability that the corresponding random variable will have a value in the interval from a to b. Also, the value of the probability density function of X at a is zero in the case of continuous random variables. The total area under the probability density function curve is equal to 1. In this thesis, we estimate the coldest temperature threshold that occurs, on average, once in N years. For each year, we are interested only in the "winter" data. Hence, we have only 91 days of data for each year. Also, we defined X_t as the random variable describing the daily average wind-adjusted temperature and T_{th} as coldest temperature threshold we want to estimate. Hence, the probability that the event $X_t \leq T_{th}$ occurs is

$$P(X_t \leq T_{th}) = \frac{1}{N * 91} \, . \tag{3.12}$$

We know the probability density function estimate and the probability that the random variable X_t has a value in the interval from $-\infty$ to T_{th}. From this information, we calculate the threshold T_{th}. For example, for $n = 67$ years of KMKE weather station daily average wind-adjusted temperatures and $N = 30$ years,

$$P(X_t \leq T_{th}) = \frac{1}{N * 91} = \frac{1}{30 * 91} = 3.663 * 10^{-4} \, . \tag{3.13}$$

Therefore, from the one-in-N algorithm and the probability in Equation 3.13, coldest

threshold temperature that occurs on average once in 30 years is $-27.1°$F using 67 years of Milwaukee daily average wind-adjusted temperatures. Similarly, we repeat this process for other weather stations in the country and for other values of N.

In this thesis, we used $N = 0.25$, 0.5, 1, 2, 5, 10, 20, and 30 years to get an idea of the trends in weather. The one-in-N threshold temperature using the one-in-N algorithm for most values of N for the weather stations KMKE, KABQ, and PANC are presented in Table 3.4. Also, we will explain more about these trends in Chapter 4.

Table 3.4: One-in-N threshold temperature (°F) using the one-in-N algorithm

N year(s)	One-in-N threshold temperature (°F)		
	KMKE	KABQ	PANC
0.25	-1.2	22.2	-2.2
0.5	-5.5	18.4	-5.5
1	-9.5	14.4	-8.7
2	-13.4	10.3	-11.8
5	-18.0	4.1	-15.4
10	-20.9	-0.2	-18.0
20	-24.5	-3.0	-21.0
30	-27.1	-4.0	-22.9

In this chapter, we reviewed the problem statement and our contribution to meteorology and statistics. We explained how the temperature data was obtained and prepared. Part of the data preparation stage was defining what "winter" means in this thesis. Then, we provided the reader with a brief description of the resulting output of the one-in-N algorithm, followed by a detailed explanation of the one-in-N algorithm. In Chapter 4, we will explain the output graphs in more detail and give examples from additional weather stations. Also, we will discuss the testing method implemented to compare the performance of the kernel density estimation method with the existing methods and evaluate the results of the tests.

CHAPTER 4

One-in-N Algorithm Test Results and Discussion of Output

4.1 Overview

In this thesis, we estimated the low temperature threshold that occurs, on average, once in N years for different weather stations in the United States using the non-parametric distribution method called kernel density estimation method. From the survey of literature, we have discussed methods used in other fields of study to estimate rare events as well as methods used in the GasDay lab to estimate the low temperature threshold. We determined that there is a possibility for improving the low temperature threshold estimate by using the kernel density estimation method. Therefore, we created the one-in-N algorithm to obtain an improved low temperature threshold estimate. In Chapter 3, we explained the one-in-N algorithm in detail. Chapter 3 also included a section on the preparation of data for use in this algorithm.

In this chapter, we will discuss the results obtained from the one-in-N algorithm for three weather stations: Milwaukee, WI (KMKE), Albuquerque, NM (KABQ), and Anchorage, AK (PANC). Then we will describe the testing method

used to show that the one-in-N algorithm is an improvement over the methods used in the GasDay lab. We will explain the results of the testing method. These results should encourage additional research in this subject.

4.2 Analysis of the One-in-N Algorithm Output

In this thesis, we use the kernel density estimation method to fit a probability density function to the daily average wind-adjusted temperature data and estimate the one-in-N low temperature threshold, where $N = 0.25$, 0.5, 1, 2, 5, 10, 20, and 30 years. So far, we have estimated the low threshold temperature. We are interested in how the estimate for the threshold temperature changes under different conditions. Specifically, we are interested in analyzing how this estimate changes over time (possibly effect of climatic changes) and with different n years of data required to calculate this estimate. Hence, we will generate two graphs for each weather station to help us analyze these trends in temperature. The first plot displays the one-in-N year conditions by n years of data used to show the effect of the length of available data. The second plot displays the one-in-N year conditions using a sliding window containing 20 years of data at a time to show the effect of a particular window of data a utility might happen to have available.

In the first graph (we name it "Increasing Window"), we are trying to evaluate whether the one-in-N conditions change over time and the minimum n

years of data needed to calculate the one-in-N low temperature threshold reliably. This plot displays the one-in-N year conditions vs. years of data used and is constructed by first using a window containing the last five years of "winter" data. We find the one-in-N low temperature threshold for the last five years and plot it on a graph. The x-axis is labeled as number of years of data used, and the y-axis is one-in-N daily average wind-adjusted temperature threshold estimate in °F. Then we find the one-in-N low temperature threshold for the last ten years and plot it on the same graph. In this way, we gradually expand the n years of data five years at a time (Hence,"Increasing Window") and estimate the one-in-N low temperature threshold for each set of n years, until the window contains all available years of "winter" temperature data. The resulting plot has eight trend lines for eight values of N. We will discuss this graph for three weather stations: KMKE, KABQ, and PANC. The discussion of the Increasing Window graph for all three weather stations will seem a little repetitive because of the overlap in the analysis of the results.

In the second graph (we name it "Sliding Window"), we try to evaluate the changes in the one-in-N low temperature threshold depending on which 20 year span we use. We also try to determine if it is prudent to obtain the low temperature threshold estimate using just 20 years of weather data or would it give us a biased estimate? This evaluation is important because some natural gas utilities estimate their design day temperature only using the coldest temperature that occurred in

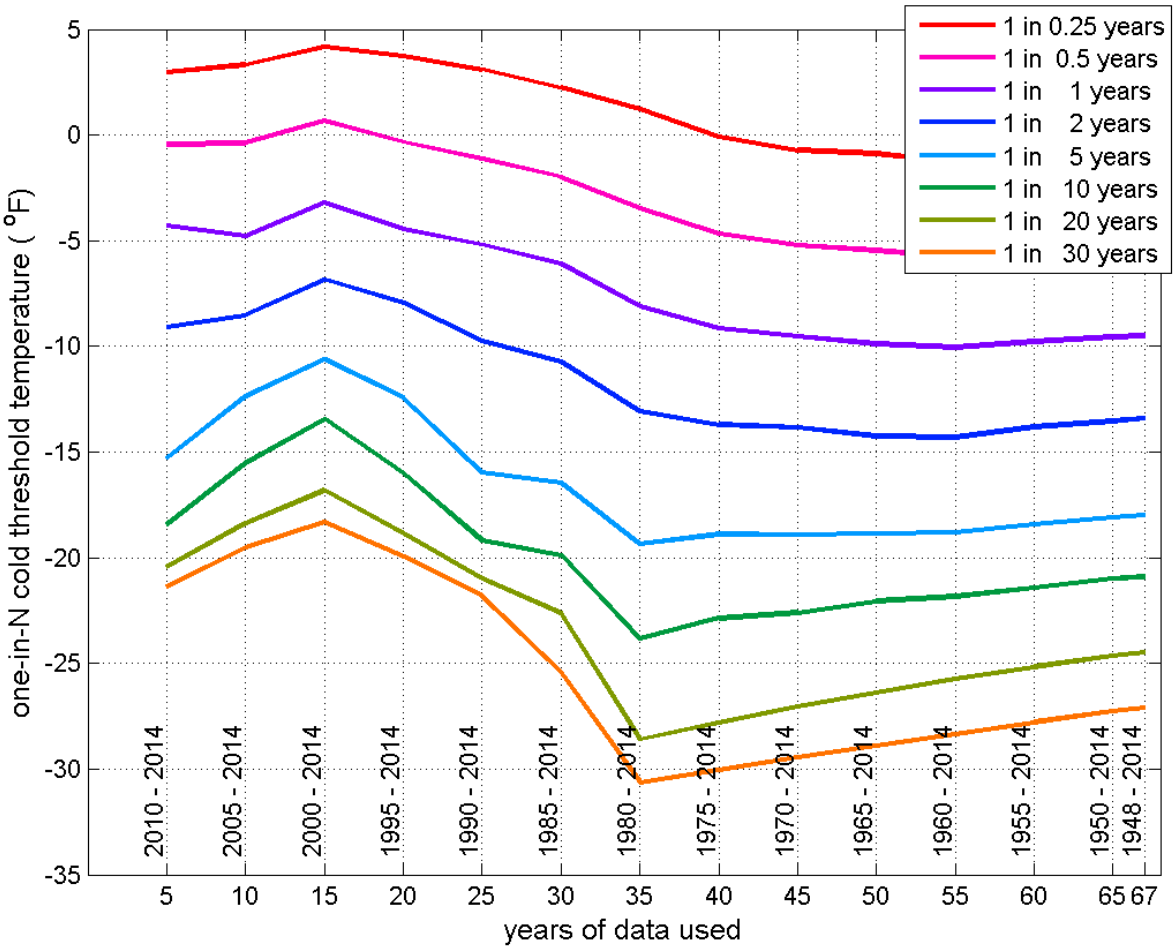


Figure 4.1: Increasing Window - Milwaukee, WI (KMKE) conditions by years of data to determine the minimum number of years needed to calculate the low temperature threshold

the past 20 or 30 years. This plot displays the one-in-N year conditions using a window containing $n = 20$ years of daily average wind-adjusted temperature. We start with the first 20 years of data, then as we add a new year of data, we drop the oldest year of data. In this manner, we "slide" this window through the data one year at a time and estimate the one-in-N year low temperature threshold after each new year is added to (and each oldest year is dropped from) the data set. As in the previous plot, we obtain a trend line for each of the N years that we are using in this thesis. As before, we will discuss this plot for the three weather stations of KMKE, KABQ, and PANC.

4.2.1 Discussion: Extreme Cold Threshold for KMKE in Milwaukee

In this section, we present an analysis of the Increasing Window (Figure 4.1) and Sliding Window (Figure 4.2) graphs constructed using the "winter" data from the weather station KMKE in Milwaukee, WI.

The purpose of the Increasing Window graph is to determine the minimum number of years needed to determine the one-in-N low temperature threshold reliably. If we look at the 1-in-30 year trend line, we can see that if we use at least 35 years of weather data, the one-in-N low temperature threshold trend line has a very small slope and is slightly increasing.

Another important aspect of this graph is instability of the rare events.

Observing all the trend lines in this graph, we can see that the rarer the event, the more unstable it is. For example, observe the difference between the 4 per year trend line and the 1-in-10 year trend line. One can see that the instability in the 1-in-10 year trend line is greater than that in the 4 per year trend line. The 1-in-30 year trend line for the low temperature threshold value from the 5-year mark to the 15-year mark shows a slightly positive slope, while this trend line shows a very distinct negative slope from the 15-year mark until the 35-year mark. These observations show that the region displaying the negative slope might be indicative of a warming climate, while the slight positive slope in the more recent years may indicate a cooling of climate, or they might be statistical fluctuations. These observations are clearer from Figure 4.2, which is explained later. However, we need many more years of data and additional research in this area to make more definitive conclusions. This gradual shift from negative to positive slopes in the trends of the low temperature threshold values is also prevalent for the other one-in-N trend lines in the graph, even though it is not as pronounced as in the rarer 1-in-30 year trend line.

Figure 4.2 is a Sliding Window graph containing a 20-year sample of weather data that slides over the entire range of available data and the one-in-N low temperature threshold is calculated with each slide. The purpose of this graph is to consider whether the rare events are stable over time. We see that the one-in-N low

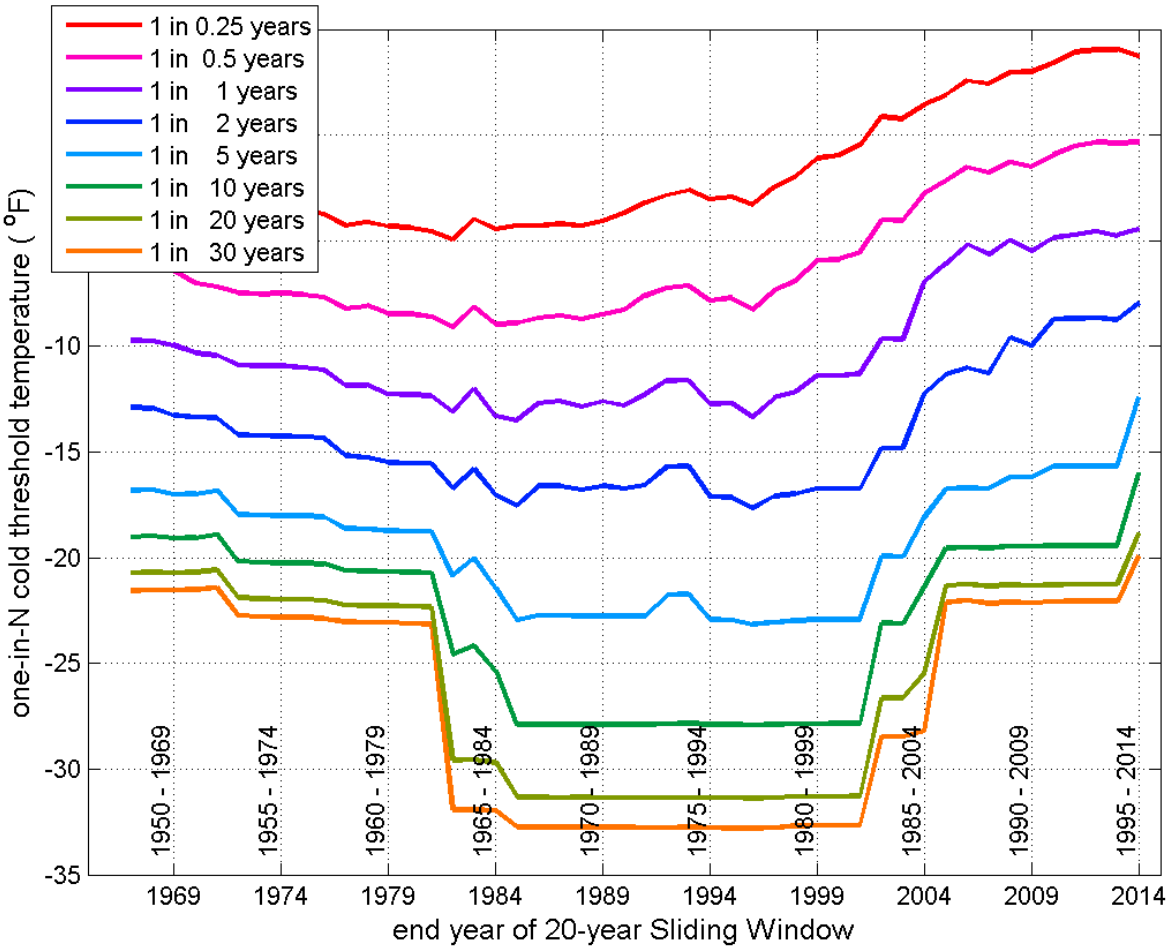


Figure 4.2: Sliding Window - Milwaukee, WI (KMKE) conditions using a window of 20 years of data to determine if conditions change with time

temperature threshold values obtained from the last 20 years of data is higher than that obtained from the first 20 years of data in the data set. For example, the 1-in-20 year trend line has a low temperature threshold value of approximately -19°F when we use the last 20 years of data, but has a low temperature threshold value of a low -21°F when we use the first 20 years of weather data from our entire

sample set. From this figure, we can see that the one-in-N algorithm is very sensitive to changes in temperature and the data in our window, particularly extremely low temperatures. In the year 1982, there was an extremely cold temperature that affected the calculation of the one-in-N low temperature threshold values (low temperature threshold values were very cold) for all the 20-year data sets that contain this year. For the subsequent data sets not containing the year 1982, the low temperature threshold values steadily increase. We see that the trend lines have a positive slope in general, which may be evidence of an overall climate warming. However, additional data and research are needed to validate this claim. We also see that the rare events are more unstable than the more frequent events, as also evident in Figure 4.1. The instability of the rare events is clearer in the graphs created for the other two weather stations explained below.

4.2.2 Discussion: Extreme Cold Threshold for KABQ in Albuquerque

In this section, we present a detailed analysis of the Increasing Window and Sliding Window graphs constructed to explain the trends in the one-in-N low temperature threshold values for Albuquerque, NM.

Figure 4.3 is the Increasing Window graph for KABQ winter data. The purpose of this graph is to determine the minimum number of years needed to explore the one-in-N low temperature threshold value. Let us observe the 1-in-10

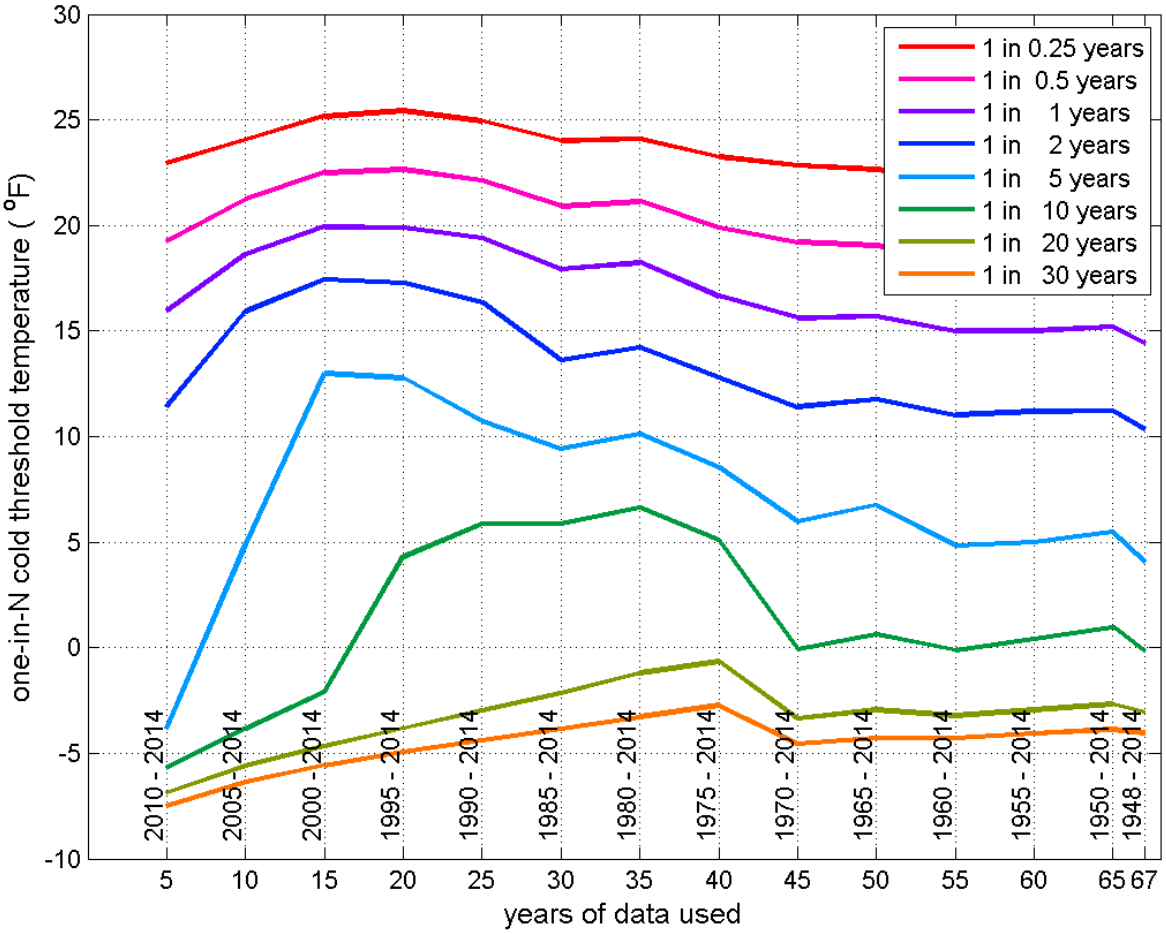


Figure 4.3: Increasing Window - Albuquerque, NM (KABQ) conditions by years of data to determine the minimum number of years needed to calculate the low temperature threshold

year trend line. On one hand, if we were to calculate the one-in-N low temperature threshold using the last 25 years of data, we would have a very high threshold temperature. On the other hand, if we calculated this threshold temperature using just the last 15 years of data, the threshold value obtained is almost 10 °F colder. Now, if we use more than 40 years of data, this one-in-N low temperature threshold value decreases along a negative slope. This may be indicative of statistical fluctuations or a slight warming of climate. However, more data and research is required to make definitive conclusions. Also, one can see that volatility of weather is more apparent for this weather station (Figure 4.3) than for Milwaukee, WI (Figure 4.1).

The Sliding Window graph (Figure 4.4) for KABQ explores whether the one-in-N low temperature threshold conditions are stable over time, using a 20-year sliding window. This is an interesting weather station to analyze because in early 2011, this area experienced an extremely rare cold event. In Figure 4.4, observe the 1-in-1 year trend line. We see that there is a general positive slope in the trend of 1-in-1 year low temperature threshold values over time. We see a general positive slope in the trend of 1-in-20 year low temperature threshold values over time as well, except for the data sets containing the year 2011. There is a sudden dip in the low temperature threshold value attributed to the extremely rare cold event. This shows that the one-in-N algorithm is very sensitive to extremes in temperature.

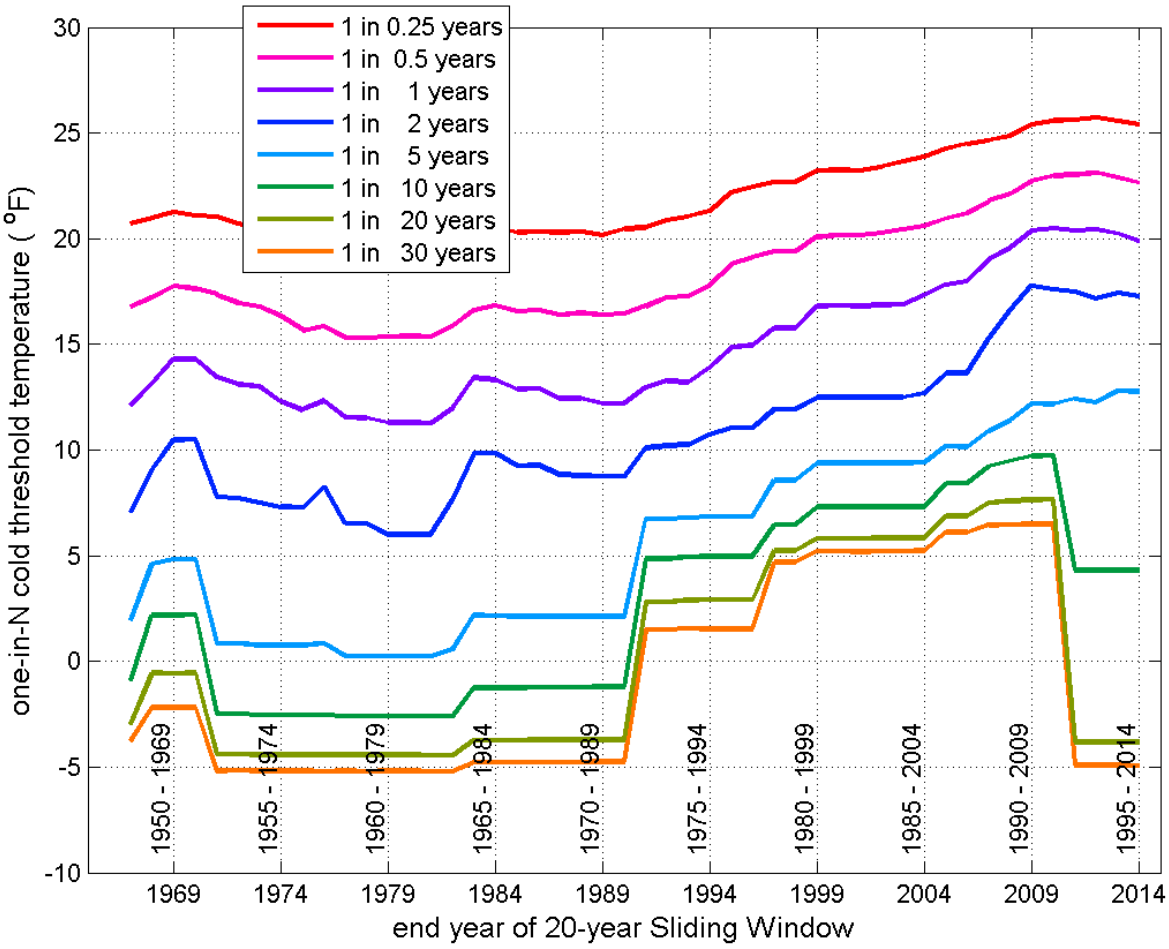


Figure 4.4: Sliding Window - Albuquerque, NM (KABQ) conditions using a window of 20 years of data to determine if conditions change with time

From this graph, we also see that volatility of weather increases as the one-in-N low temperature threshold values get rarer. However, in general, the Albuquerque winter seems to be getting warmer as the trend lines seem to have an overall positive slope. In particular, the extreme cold temperature that occurred on February 2^{nd}, 2011, shows that the weather is quite volatile for rare events, even

though the general trend of temperature may be warmer. Hence, LDCs should not ignore the rarest one-in-N low temperature threshold simply because there may be a general warming trend.

4.2.3 Discussion: Extreme Cold Threshold for PANC in Anchorage

In this section, we will analyze the Increasing Window and the Sliding Window graphs for the weather station in Anchorage, AK (PANC). This weather station was chosen for analysis because in 1989 they experienced their low temperature threshold conditions on three consecutive days.

From Figure 4.5, we are trying to consider how many years are sufficient to calculate the one-in-N low temperature threshold values. Analyzing this figure, we see that the one-in-N low temperature threshold values have a near zero slope when we use more than 30 years of data. Looking at the 1-in-30-year trend line, we see that there has not been a low temperature threshold condition since 1989, which explains the higher values for the 1-in-30-year low temperature threshold calculations for windows of data that do not contain the year 1989. This graph is a good example to show instability of the rare events compared to the more frequent events. The trend lines below the 1-in-2-year line show increasing instability, which is obvious around the 15 year mark, where it displays a sudden dip in temperature.

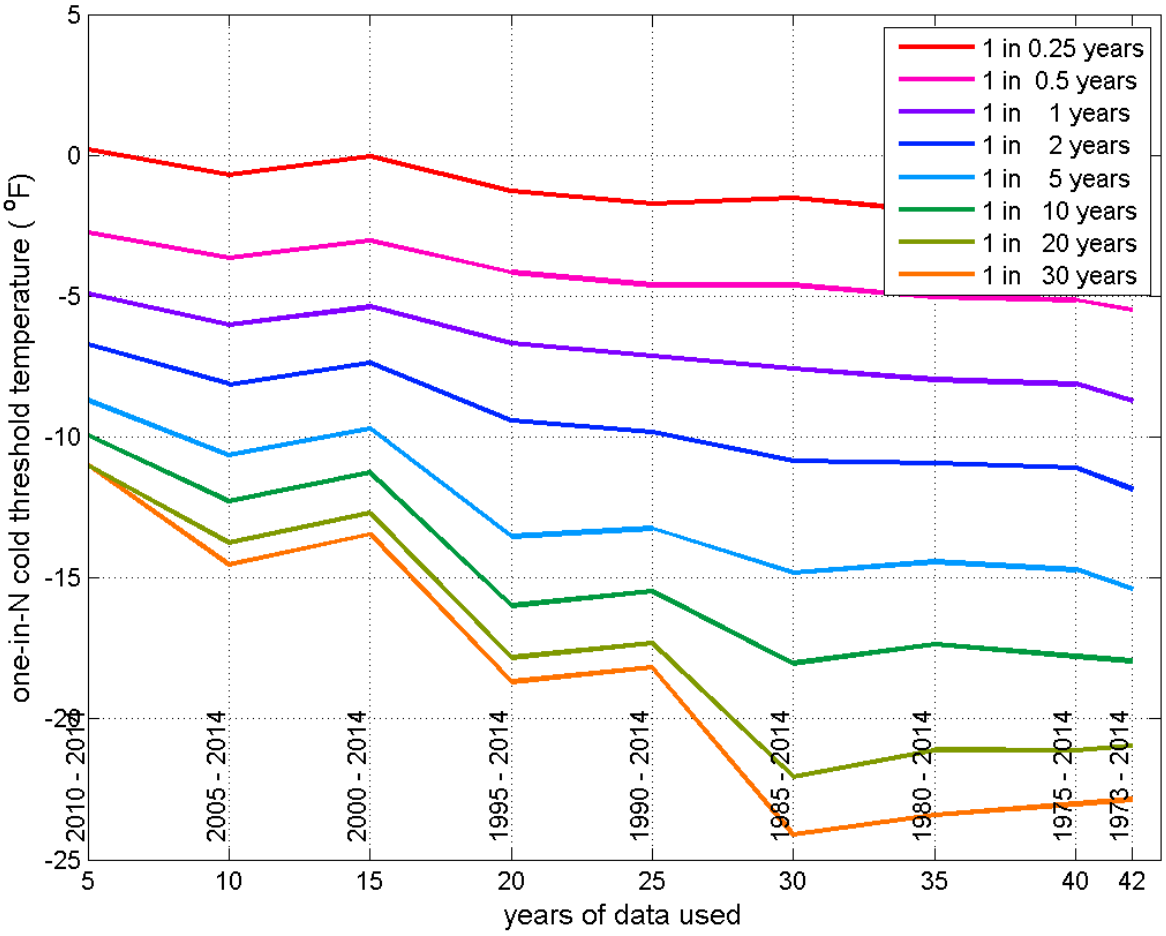


Figure 4.5: Increasing Window - Anchorage, AK (PANC) conditions by years of data to determine the minimum number of years needed to calculate the low temperature threshold

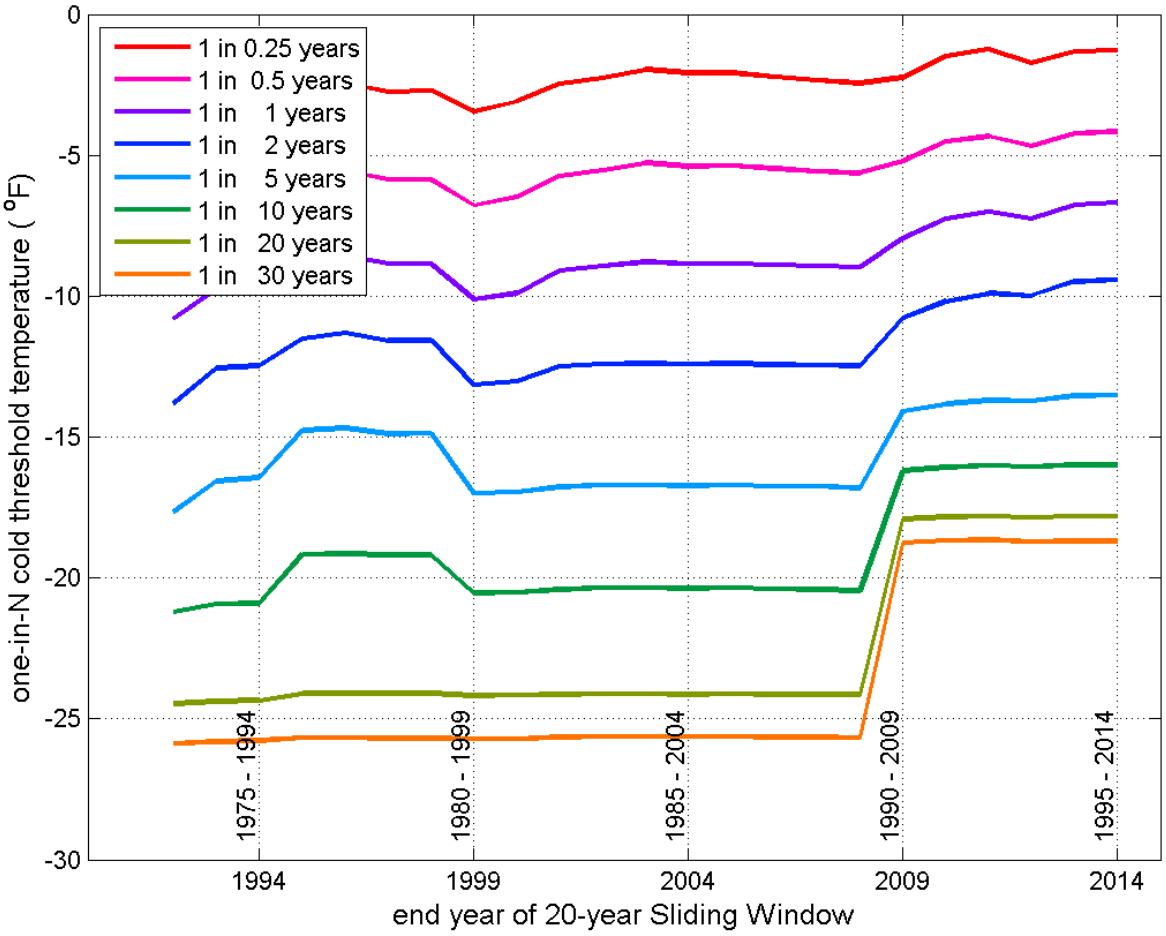


Figure 4.6: Sliding Window - Anchorage, AK (PANC) conditions using a window of 20 years of data to determine if conditions change with time

Figure 4.6 helps us explores whether the one-in-N low temperature threshold conditions are changing over time. Here, we see that the three extremely cold consecutive days in 1989 significantly affect the calculation of the 1-in-20 year and 1-in-30 year low temperature threshold values. When 1989 falls out of the 20 year window, these low temperature threshold values become significantly warmer (by

approximately 5 or 6 °F). This observation confirms that the one-in-N algorithm is very sensitive to extreme cold temperatures and the size of the data set. It also shows that as the weather becomes increasingly volatile, the rarer the events get.

From this graph, there does not seem to be an obvious slope in the trend lines, indicating that over the available history, the winter weather has remained nearly constant. In other words, for this weather station, the climate doesn't seem to have warmed or cooled or has not shown much statistical fluctuation. However, more data and research are required to make definitive conclusions about these preliminary observations.

4.3 Test and Results

So far in this thesis, we explained the One-in-N Algorithm, showed that it models the winter daily average wind-adjusted temperature better than the methods currently used in the Gasday lab at the 5% statistical level of significance, and discussed the results obtained. Now, we show that the one-in-N low temperature threshold estimates from the One-in-N Algorithm is more accurate compared to the one-in-N low temperature threshold estimates from the generalized extreme value distribution, normal distribution, and the variance weighted composite method.

1. We separate the n years of daily average wind-adjusted temperature data into a training set and a test set. The training set comprises of 30 years of winter daily average wind-adjusted temperature data selected at random (pick 30 years from available n years randomly, and each chosen year contains the entire winter data for that year only). The test set comprises of the remaining $n - 30$ years of daily average wind-adjusted temperature data.

2. Using the training set, we calculate one-in-N low temperature threshold estimates, where $N = 30$ years, using the one-in-N algorithm, generalized extreme value distribution, normal distribution, and the variance weighted composite. Then we count the number of times that the daily average wind-adjusted temperature from the test set is less than or equal to these four one-in-30 year estimates (called `testcount`).

3. We repeat steps 1 and 2 a hundred times and compare the mean count of the daily average wind-adjusted temperature from the test set less than or equal to the one-in-N low temperature threshold estimates (called `meantestcount`), obtained from the generalized extreme value distribution, the normal distribution, the variance-weighted composite distribution, and the kernel density estimate to the theoretical expected count

$$\texttt{theoreticalexpectedcount} = \frac{\text{size of test set}}{30} . \tag{4.1}$$

We performed this test using the data from KMKE, KABQ, and PANC weather stations.The results of this test, for each weather station, are shown in Table 4.1.

Table 4.1: Validation test results after 100 iterations

Methods	Mean one-in-30 year threshold temperature (°F)		
	KMKE	KABQ	PANC
one-in-N algorithm	−26.5190	−3.1770	−22.8081
generalized extreme value distribution	−17.4860	7.9787	−24.5158
normal distribution	−17.4602	10.0613	−19.1824
variance weighted composite method	−17.5432	9.0314	−22.7026

Methods	mean count of test-set data $\leq$ mean one-in-30 year estimate		
	KMKE	KABQ	PANC
theoretical expected	1.2333	1.2000	0.3667
one-in-N algorithm	1.7700	2.3600	0.5200
generalized extreme value distribution	9.0100	12.6200	0.3300
normal distribution	8.7500	17.2800	0.6000
variance-weighted composite method	8.6300	14.3900	0.5100

We need to determine if the `meantestcount` is statistically different from the `theoreticalexpectedcount`. For each of the generalized extreme value

distribution, the normal distribution, the variance-weighted composite distribution, and the kernel density estimate, we calculate the square of the difference between each `testcount` and the `theoreticalexpectedcount` (called square error). Then we perform a t-test of statistical significance. We find that for KMKE and KABQ, the `meantestcount` is statistically different from the `theoreticalexpectedcount` at the 5% level. However, for PANC, we are unable to conclude that the `meantestcount` is statistically different from the `theoreticalexpectedcount`.

These results demonstrate that the low temperature threshold estimate from the one-in-N algorithm is more accurate than the low temperature threshold estimates from the generalized extreme value distribution, normal distribution, and the variance weighted composite method for KMKE and KABQ. In the future, we should explore increasing the winter window for PANC to see if we can obtain a `meantestcount` that is statistically different from the `theoreticalexpectedcount`.

CHAPTER 5

Conclusions and Future Research

5.1 Conclusions

Our goal was to develop an algorithm to estimate the one-in-N low temperature threshold value for weather stations in the United States that was better than the existing methods. By applying statistics to the weather, particularly nonparametric methods, we developed the one-in-N algorithm to estimate the one-in-N low temperature threshold value.

We explained how the one-in-N algorithm was developed and used in Chapter 3. In Chapter 4, we showed how this algorithm is better than the current methods used. We used the RMSE measure to compare the fit of data to the existing methods with the fit of data to the kernel density estimation method and found that the kernel density estimation method provided a better fit to the temperature distribution data. The kernel density estimation method is a major part of the one-in-N algorithm. Then we created a test scenario using a training set and a test set of weather data per station and evaluated the number of times we obtained or deceded the one-in-N low temperature threshold. We found that the

one-in-N algorithm provided a more reliable estimate than the existing methods. From the explanation of the kernel density estimation method in Chapter 2, we know that the density estimation near a point consists of contributions from kernels above and below that point. However, for the minimum value of the observation (let us call it XMIN), we cannot compute the kernel contributions below this point because we do not have that data. If we only use the kernels above XMIN, it will make a biased density estimate. To reduce this bias, we compute the contributions from kernels centered above XMIN, and fold their values around XMIN. The result should be good if the density is nearly flat in this area. If the density is increasing, then the estimate will still be biased downward, and if the density is decreasing, it will still be biased upward, but the bias will be reduced.

5.2 Future Research

In this thesis, we investigated and developed a method to estimate the one-in-N low temperature threshold value. We also used this method to try to answer two additional questions:

1. How many years of data is needed to reliably estimate the one-in-N low temperature threshold?
2. Are these conditions changing over time?

However, there are still many improvements and extensions that can be applied to this method. We list a few suggestions to improve the one-in-N method, and a few hypotheses we can test that were suggested by other scientists.

5.2.1 Extension of Work

In this thesis, we detrended the data using a time-varying mean. An extension to this method could be to develop a technique to detrend the temperature data with a time-varying variance, skewness, and kurtosis. The intention is that at the end of this process, we will be left with only the pure, unaltered, underlying data, which means that location and season would not be a restriction anymore. Hence, we could use all the available weather data from all the weather stations, and for all seasons, resulting in a really large data set. The kernel density estimation method works better for large amounts of data.

Another extension to this algorithm could be trying to find the one-in-N low temperature threshold per month. This approach may give one a better idea of when these extreme cold events occur in a particular month and perhaps provide insight into whether there is a distinct pattern in the occurrences per month for each weather station.

For LDCs, knowing the one-in-N low temperature threshold for the day is

only partially helpful to plan the day's natural gas needs. We could extend this thesis work to find the one-in-N low temperature threshold hour, which would help identify the hour in a day that would require the most natural gas use.

In this thesis, we used a fixed 91 days with the coldest daily average wind-adjusted temperatures to find the one-in-N low temperature threshold. Our experiments showed that this 91 days with the coldest daily average wind-adjusted temperatures was sufficient to contain the high variance regions of data, namely, winter. However, for certain weather stations, we could expand this window of data, while for others, we could narrow this window of data. Hence, one could develop an algorithm to use a variable window of data that expands or contracts to the high variance regions in the data.

From Chapter 2, Mearns [6] identified several characteristics of the time-series data; one in particular, the autocorrelation function, which is a measure of the dependence among the data points. We can extend this work by identifying the autocorrelation between the data points and divising a solution to handle autocorrelation in the data.

We also could study how to handle time of observation changes, station changepoints (location/instrument), and urbanization, introduced by Knappenberger [11], as they affect data quality and analysis of trends in temperature.

5.2.2 Hypotheses to be Explored

According to Meehl [2] and Mearns [6], a small change in the mean temperature causes shifts of practical significance in the probabilities of extreme temperature events. More research can be applied to locations such as Albuquerque, which experienced a rare cold temperature event in 2011, to discern if there was any noticeable change in the mean or variance from prior years.

Another hypothesis suggested by Meehl [7] is that variability in summer has increased, and variability in winter has decreased. Lu [9] found that winters show most warming in the northern MidWest, the four corners region, the Dakotas, south Arizona, and southern California. Also, the winters seem to be warming more than the summers. We could extend our research to include summer months to determine whether we can confirm this hypothesis.

Easterling [13], Knappenberger [11], and Lu [9] found that night-time temperatures have been increasing more than the day-time temperatures. We could determine whether we can make similar conclusions using our one-in-N algorithm and explore how this change affects the calculation of the design day conditions. Besides, an increased ability to monitor and detect multidecadal variations and trends is critical to detect changes in trends and to understand their origins [13].

Following from the item above, we could explore the hypothesis that extreme

cold events change with location and elevation, as suggested by Rebetez [14]. We could also investigate the effect of the North Atlantic Oscillation (NAO) index on extreme cold events.

Further research can be made to estimate the one-in-N low temperature threshold in the rest of the world, provided that sufficient data is available. A part of this initiative could be to analyze the effect of El Niño Southern Oscillations (ENSO) on the predictability of extreme cold temperature frequency. According to Gershunov [8], weak ENSO events improve the predictability of extreme cold temperature frequency in the South United States. Also, weak La Niña winters improve the extreme cold temperature frequency predictability in the Midwest. Future research into ENSO events, may provide additional insight into how the one-in-N low temperature threshold changes with time and how it is correlated with the type of winter (El Niño or La Niña).

In conclusion, we investigated, researched, and developed the one-in-N algorithm to estimate the one-in-N low temperature threshold value using a non-parametric distribution called the kernel density estimation method. We compared the output of the one-in-N algorithm with the outputs of the generalized extreme value distribution method, the normal distribution method, and the variance-weighted distribution method. We validated that the one-in-N algorithm provides a better estimate for the one-in-N low temperature threshold at the 5%

level of significance. We also used this method to analyze two questions; how many years of data is needed to estimate the one-in-N low temperature threshold accurately, and are these conditions changing over time? We provided a few suggestions to improve the one-in-N method and a few hypotheses suggested by other scientists to be explored.

BIBLIOGRAPHY

[1] P. Potocnik, *Natural Gas*. Rijeka, Croatia: Sciyo, August 2010, www.sciyo.com; http://www.intechopen.com/books/show/title/natural-gas.

[2] G. A. Meehl, T. Karl, D. R. Easterling, S. Changnon, R. Pielke, D. Changnon, J. Evans, P. Y. Groisman, T. R. Knutson, K. E. Kunkel, L. O. Mearns, C. Parmesan, R. Pulwarty, T. Root, R. T. Sylves, P. Whetton, and F. Zwiers, "An introduction to trends in extreme weather and climate events: Observations, socioeconomic impacts, terrestrial ecological impacts, and model projections*," *Bulletin of the American Meteorological Society*, vol. 81, no. 3, pp. 413–416, 2000.

[3] W. Hardle, M. Muller, S. Sperlich, and A. Werwatz, *Nonparametric and Semiparametric Models: An Introduction*. New York: Springer, 2004.

[4] J. Fahey, "Natural gas soars as cold grips nation," *USA Today*, January 24, 2014, [Online]. Available: http://www.usatoday.com/story/money/business/2014/01/24/natural-gas-soars-as-cold-grips-homes-drillers/4843033/. [Accessed Feb. 24, 2014].

[5] N. M. Public Regulation Commission informal task force investigation, "Severe weather event of February, 2011 and its cascading impacts on NM utility service," December 21, 2011, [Online]. Available: http://www.nmprc.state.nm.us/utilities/docs/2011-12-21_Final_Report_NMPRC.pdf. [Accessed Feb. 26, 2014].

[6] L. O. Mearns, R. W. Katz, and S. H. Schneider, "Extreme high-temperature events: Changes in their probabilities with changes in mean temperature," *Journal of Climate and Applied Meteorology*, vol. 23, no. 12, pp. 1601–1613, 1984.

[7] G. A. Meehl, F. Zwiers, J. Evans, T. Knutson, L. Mearns, and P. Whetton, "Trends in extreme weather and climate events: Issues related to modeling extremes in projections of future climate change," *Bulletin of the American Meteorological Society*, vol. 81, no. 3, pp. 427–436, 2000.

[8] A. Gershunov, "ENSO influence on intraseasonal extreme rainfall and temperature frequencies in the contiguous United States: Implications for long-range predictability," *Journal of Climate*, vol. 11, no. 12, pp. 3192–3203, 1998.

[9] Q. Q. Lu, R. Lund, and L. Seymour, "An update of U.S. temperature trends," *Journal of Climate*, vol. 18, no. 22, pp. 4906–4914, 2005.

[10] S. J. Armstrong, *Principles of Forecasting a Handbook for Researchers and Practitioners*. Norwell, MA: Kluwer Academic Publishers, 2001.

[11] P. Knappenberger, P. Michaels, and R. Davis, "Nature of observed temperature changes across the United States during the 20th century," *Climate Research*, vol. 17, pp. 45–53, 2001.

[12] M. Menne, R. Vose, C. Williams, Jr., and D. Easterling, "United States Historical Climatology Network daily temperature, precipitation, and snow data," *Database, The Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy*, 2006.

[13] D. Easterling, J. L. Evans, P. Y. Groisman, T. R. Karl, K. E. Kunkel, and P. Ambenje, "Observed variability and trends in extreme climate events: A brief review," *Bulletin of the American Meteorological Society*, vol. 81, no. 3, pp. 417–425, 2000.

[14] M. Rebetez, "Changes in daily and nightly day-to-day temperature variability during the twentieth century for two stations in Switzerland," *Theoretical and Applied Climatology*, vol. 69, no. 1-2, pp. 13–21, 2001.

[15] I. Bell, "USDA plant hardiness zone map," September 13 2011, http://www.ldeo.columbia.edu/res/pi/NAO/.

[16] S. M. Robeson, "Trends in time-varying percentiles of daily minimum and maximum temperature over North America," *Geophysical Research Letters*, vol. 31, no. 4, 2004.

[17] F. K. Lyness, "Consistent forecasting of severe winter gas demand," *The Journal of the Operational Research Society*, vol. 32, no. 5, pp. 347–359, 1981. [Online]. Available: http://www.jstor.org/stable/2581552

[18] S. Jaimungal and E. Ng, *Kernel-Based Copula Processes*, ser. Machine Learning and Knowledge Discovery in Databases. Springer Berlin / Heidelberg, 2009, vol. 5781, pp. 628–643.

[19] P. M. Dixon, A. M. Ellison, and N. J. Gotelli, "Improving the precision of estimates of the frequency of rare events," *Ecology*, vol. 86, no. 5, pp. 1114–1123, 2005.

[20] S. Babik, "Application of reliability theory to a reactor safety circuit," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 17, no. 2, pp. 137–156, 1968. [Online]. Available: http://www.jstor.org/stable/2985677

[21] W. A. Thompson Jr., "Some reliability models for safety assessment," *Operations Research*, vol. 32, no. 3, Reliability and Maintainability, pp. 559–565, May - Jun. 1984. [Online]. Available: http://www.jstor.org/stable/170474

[22] H. M. Cathey, "USDA plant hardiness zone map," October 6 2003, http://www.usna.usda.gov/Hardzone/hrdzon2.html.

[23] J. L. Devore, *Probability and Statistics for Engineering and the Sciences - Fourth Edition*. Belmont, CA: International Thomson Publishing Company, 1995.

[24] E. Weisstein, "Wolfram mathworld," September 2011, http://mathworld.wolfram.com/topics/ContinuousDistributions.html.

[25] S. Kotz and S. Nadarajah, *Extreme Value Distributions: Theory and Applications*. London; River Edge, NJ: Imperial College Press, 2000.

[26] P. Embrechts, C. Kluppelberg, and T. Mikosch, *Modelling Extremal Events for Insurance and Finance*, ser. Applications of mathematics. Berlin; New York: Springer, 1997, vol. 33.

[27] M. He, "Annual transformation algorithm for hourly data in natural gas consumption," *Marquette University Thesis (M.S.)*, 2007.

[28] "American Gas Association," 2014, http://www.aga.org.

[29] S. Combs, "The energy report 2008," *Texas Comptroller of Public Accounts*, vol. 96, no. 1266, p. 72, 2008. [Online]. Available: http://www.window.state.tx.us/specialrpt/energy/

[30] J. E. Freund, I. Miller, and M. Miller, *Mathematical Statistics with Applications*, 7th ed. New Jersey: Pearson, October 2003.

[31] J. D. Gibbons, *Nonparametric Statistics: An Introduction*. Newbury Park, CA: Sage Publications, 1993, vol. 90.

[32] P. Glasserman, P. Heidelberger, P. Shahabuddin, and T. Zajic, "Multilevel splitting for estimating rare event probabilities," *Operations Research*, vol. 47, no. 4, pp. 585–600, Jul. - Aug. 1999, http://www.jstor.org/stable/223163.

[33] S. Jackman, "Maximum likelihood," p. 24, March 14 2006, lecture from Stanford University, http://jackman.stanford.edu/classes/350B/06/maxlik.pdf.

[34] J. Kantor, "What's the temperature supposed to be?" 2010, private communication.

[35] A. Papoulis and U. S. Pillai, *Probability, Random Variables and Stochastic Processes*. New York: McGraw-Hill, 2001.

[36] A. M. Polovko, *Fundamentals of Reliability Theory*. New York: Academic Press, 1968.

[37] S. Siegel, "Nonparametric statistics," *The American Statistician*, vol. 11, no. 3, pp. 13–19, Jun. 1957. [Online]. Available: http://www.jstor.org/stable/2685679