# Improving Gas Demand Forecast During Extreme Cold Events

Babatunde Isaac Ishola
*Marquette University*

# IMPROVING GAS DEMAND FORECAST DURING EXTREME COLD EVENTS

by

Babatunde I. Ishola, B.S.

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

Milwaukee, Wisconsin

December 2016

## ABSTRACT

IMPROVING GAS DEMAND FORECAST DURING EXTREME COLD EVENTS

Babatunde I. Ishola, B.S.

Marquette University, 2016

This thesis explores techniques by which the accuracy of gas demand forecasts can be improved during extreme cold events. Extreme cold events in natural gas demand data are associated with large forecast error, which represents high business risk to gas distribution utilities.

This work begins by showing patterns associated with extreme cold events observed in natural gas demand data. We present a temporal pattern identification algorithm that identifies extreme cold events in the data. Using a combination of phase space reconstruction and a nearest neighbor classifier, we identify events with dynamics similar to those of an observed extreme event. Results obtained show that our identification algorithm (RPS-$k$NN) is able to successfully identify extreme cold events in natural gas demand data.

Upon identifying the extreme cold events in the data, we attempt to learn the residuals of the gas demand forecast estimated by a base-line model during extreme cold events. The base-line model overforecasts days before and underforecasts days after the coldest day in an extreme cold event due to an unusual response in gas demand to extreme low temperatures. We present an adjustment model architecture that learns the pattern of the forecast residuals and predicts future values of the residuals. The forecasted residuals are used to adjust the initial base model's estimate to derive a new estimate of the daily gas demand. Results show that the adjustment model only improves the forecast in some instances.

Next, we present another technique to improve the accuracy of gas demand forecast during extreme cold events. We begin by introducing the Prior Day Weather Sensitivity (PDWS), an indicator that quantifies the impact of prior day temperature on daily gas demand. By investigating the complex relationship between prior day temperature and daily gas demand, we derived a PDWS function that suggests PDWS varies by temperature and temperature changes. We show that by accounting for this PDWS function in a gas demand model, we obtain a gas model with better predictive power. We present results that show improved accuracy for most unusual day types.

# ACKNOWLEDGMENTS

Babatunde I. Ishola, B.S.

I would like to thank the GasDay Lab at Marquette University for providing me the platform and support which enabled me to undertake this research work.

I am especially grateful to my supervisor, Dr. Ronald Brown for providing me with this unique opportunity, providing insights and direction, and for believing in my ability. A special thanks to Dr. George Corliss for his guidance and support throughout this work. The scientific input from him has proven to be of tremendous value in every area of this thesis. I would also like to express my gratitude to Dr. Richard Povinelli for his contributions in Chapters 2 and 3 of this thesis.

A special thanks all the members of the GasDay Lab. Notably, Andrew Kirkham for his willingness to read and offer suggestions to improve this publication. Jarrett Smalley, for assisting in research effort of Chapter 4. Other past and present graduate and undergraduate students of the GasDay lab, including but not limited to, Paul Kaefer, William Castedo, Nick Winninger, Hermine Akouemo, Maral Fakoor, Andrew Tran, Greg Merkel, Mohammad Saber, and David Kaftan for their support all the way.

Many thanks to my thesis committee in the persons of Dr. Ronald Brown, Dr. George Corliss, and Dr. Michael Johnson for their valuable advise, comments, and scrutiny.

This work is dedicated to my mother, Ruth and brothers Samuel and Emmanuel.

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

# CHAPTER 1

# Natural Gas Demand Forecasting

Natural gas demand forecasting involves predicting future values of gas demand based on observations of historical gas consumption and its predictor variables. Natural gas distribution companies (utilities) use the predicted values in their decision-making process. Hence, they need accurate forecasts. This work aims to help gas utilities by considering methods by which the accuracy of gas demand forecasts can be improved, especially during periods of extreme cold weather.

In this chapter, we provide background information on natural gas and its uses, highlight the importance of forecasting gas demand accurately, and point out how improvement in accuracy impacts the gas utilities. The specific problems this thesis addresses are introduced. Finally, the organization map of this thesis is presented.

## 1.1 Introduction to Natural Gas

Natural gas, being the cleanest and most efficient fossil fuel, is an essential part of the United States energy industry [12]. Natural gas supplies nearly one-fourth of the energy used in the United States [3]. Natural gas is an abundant resource across the

United States, but it is a non-renewable resource. Most of the natural gas consumed in the U.S. is produced domestically and distributed to the end users by gas utilities [64]. The utilities are responsible for getting natural gas from production and distributing it to consumers for residential, commercial, and industrial usage. This is achieved through massive underground pipeline distribution systems spanning more than two million miles [43].

Natural gas is an intrinsic part of the nation's energy supply. Industries, the largest consumer of natural gas, rely on natural gas for power plants; for electric load generation; for use as base compound for chemicals and as a manufacturing feed stock for products such as plastic, fabrics, fertilizer, and pharmaceutical products; for waste treatment, metal preheating, and industrial boilers [44]. The commercial users of natural gas includes offices, schools, hotels, churches, hospitals, and government buildings [44]. In the commercial sector, natural gas is mostly used for space heating, cooling, cooking, and water heating. Residential consumption of natural gas is similar to commercial use. Residential users are mainly homes, and most of natural gas is used in homes for space heating [44]. According to the U.S. Energy Information Administration (EIA), space heating accounts for 65% of residential usage in 2014, with up to 56 million households using natural gas for space heating [65].

## 1.2 Importance of Accurate Forecasting

For gas utilities to meet the customer's gas demand effectively, the utilities must forecast both the short-term and long-term demand for gas. Short-term forecasts are important for daily or weekly plans [4, 19], while long-term forecasts help in making design and long-term plans such as making sure their infrastructure is capable of handling future expected high demand and having enough gas for distribution [37, 40]. In both cases, forecasting gas demand accurately is extremely important to fulfill the end customers' gas demand.

Gas utilities themselves buy natural gas from gas suppliers, usually by nominating in advance. The nomination is based on the projected end customers' demand. The utilities are penalized for under-forecasting or over-forecasting the nomination, which could translate to millions of dollars in loss for the utilities and/or increased purchase cost for the consumers. For example, if the utilities under-forecast the demand, they may be forced buy extra gas at very high spot market prices. On the other hand, if they over-forecast the demand and nominate more gas than needed, the utility is penalized for not taking the agreed amount of gas from the transmission lines. Storing up the extra gas is not always a viable option, as they also incur extra cost for managing addition storage facilities. These and other operational factors discussed later in this thesis motivate accurate forecasting of gas demand.

## 1.3 Uses of Natural Gas

In forecasting the amount of natural gas consumers will require each day, it is important to understand the nature of the end customers' gas consumption, which depends on their specific use for natural gas. Industrial use is often fairly constant or influenced by the industry's product demand. Industrial consumption usually is temperature independent. Residential and commercial use is influenced by current weather conditions, as most gas is used for space heating. Residential and commercial gas demand is driven by weather factors such as temperature, wind, dew point, and cloud cover. Of theses factors, temperature is the major determining factor.

Daily temperature is the most significant factor influencing the day's demand. Gas demand increases as the temperature decreases. Homes and businesses use more natural gas for space heating as it gets colder. The highest gas demand occurs during the winter (heating season), while the lowest gas demand occurs during the summer, when no space heating is required. Figure 1.1 shows typical daily gas consumption data. The data presented in Figure 1.1 are the actual natural gas consumption data over a period of ten years obtained from a gas utility in the United States, scaled to protect the identity of the utility. The periodic spikes in the gas consumption data reflect the high gas demand during the winter. In Figure 1.2, the daily gas consumption (flow) data is plotted against the

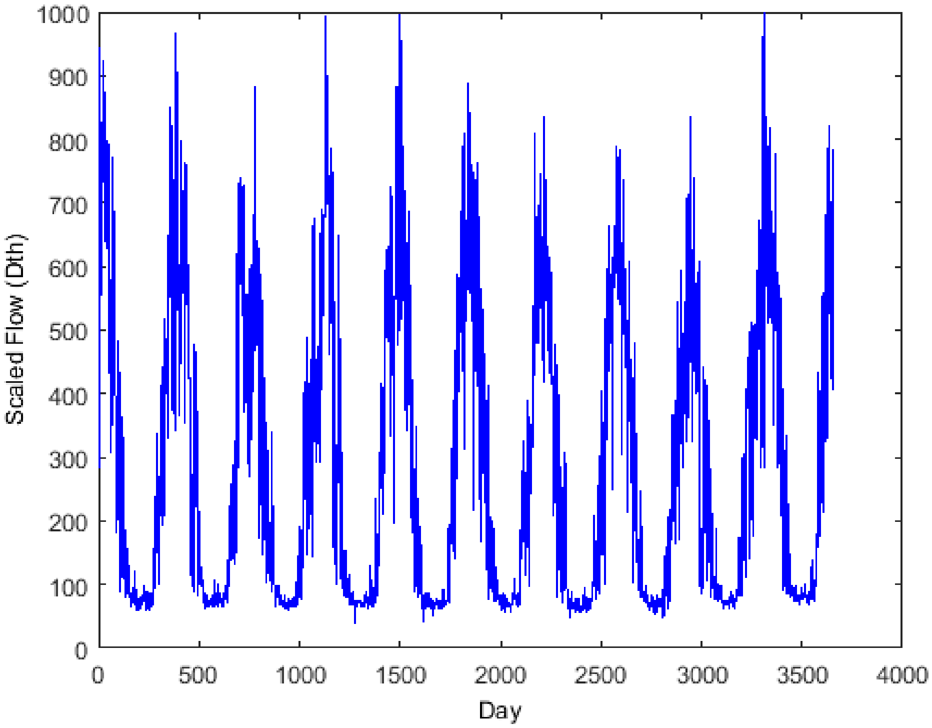


Figure 1.1: Gas consumption data over a 10 year period

corresponding daily temperature showing the almost-linear relationship between gas demand and temperature.

## 1.4 Organization of Thesis

This thesis a sum of efforts towards improving the accuracy of natural gas demand forecasts during periods of extreme cold events. The work done is presented in Chapters 2, 3, and 4 as three independent documents. Each is a self-sufficient document having its own introduction, background, method, discussion, and conclusion. Each chapter either builds on the previous chapter or explores a

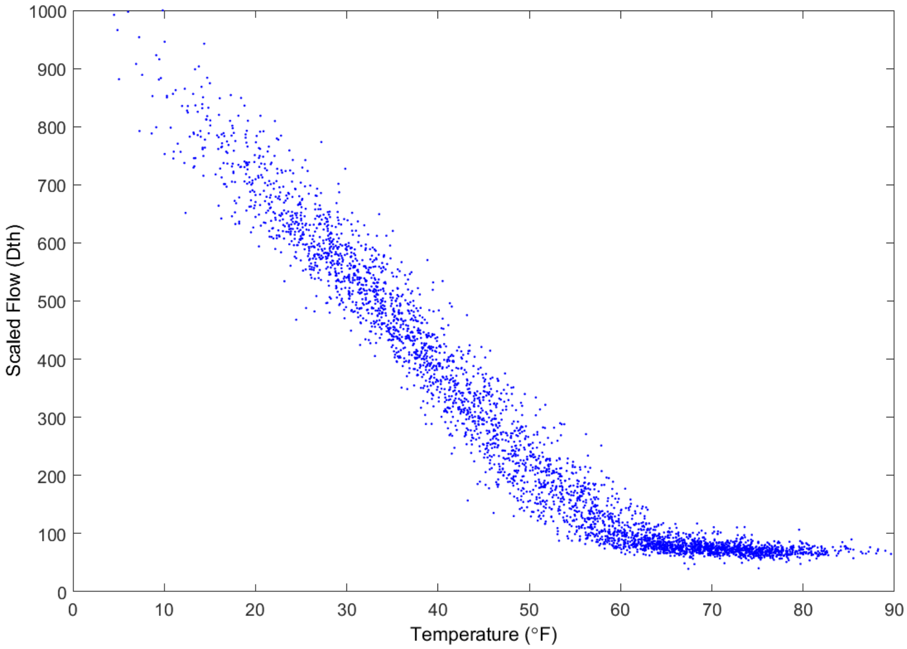


Figure 1.2: Gas flow against temperature

different method. For the three main chapters, the overall objective is to achieve improved forecast during extreme cold events. General introductory and concluding sections are presented in Chapters 1 and 5. A visual layout of the organization of this thesis is provided in Figure 1.3.

Chapter 2 discusses identifying extreme cold events in natural gas data using a temporal pattern identification algorithm. Periods of extreme cold events are characterized by dynamics different from most common days. We present a semi-supervised identification algorithm that clusters events based on similarity. Using a combination of a phase space reconstruction technique and a nearest

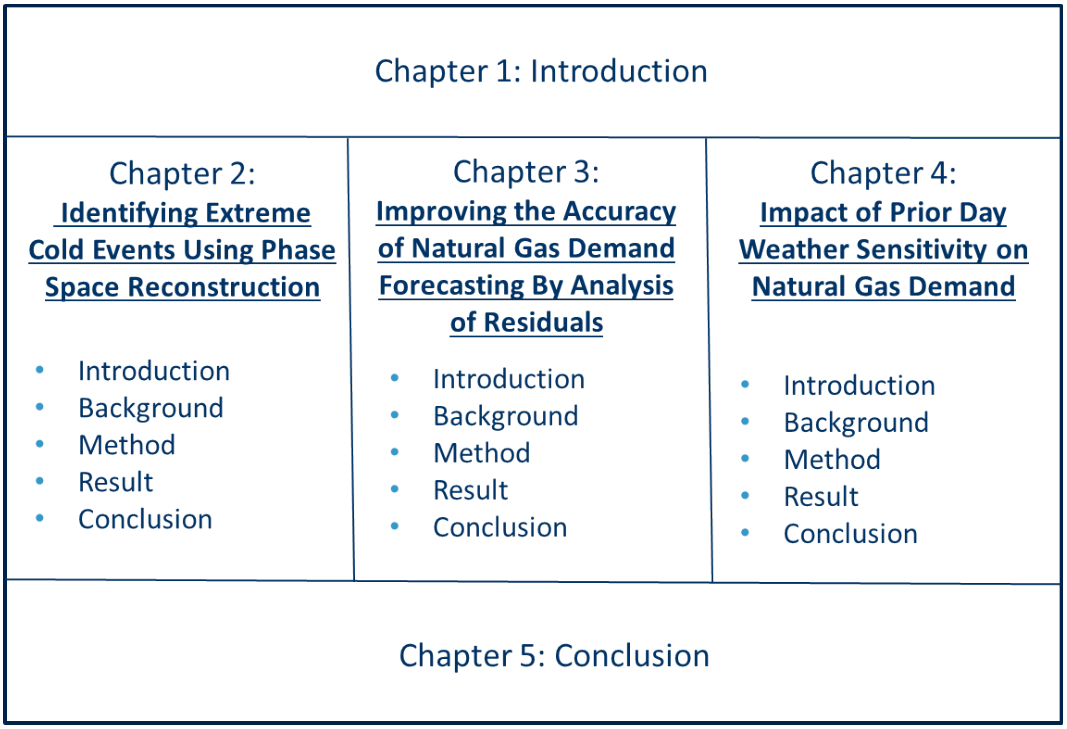

Chapter 1: Introduction
Chapter 2:
Identifying Extreme Cold Events Using Phase Space Reconstruction
Introduction
Background
Method
Result
Conclusion
Chapter 3:
Improving the Accuracy of Natural Gas Demand Forecasting By Analysis of Residuals
Introduction
Background
Method
Result
Conclusion
Chapter 4:
Impact of Prior Day Weather Sensitivity on Natural Gas Demand
Introduction
Background
Method
Result
Conclusion
Chapter 5: Conclusion

Figure 1.3: Thesis layout

neighbour algorithm, we identify other extreme cold events similar to an observed extreme cold event in the data.

Chapter 3 is an extension of the work in Chapter 2. Chapter 3 discusses how the extreme cold events identified in Chapter 2 can be used to improve the accuracy of gas demand forecast during extreme cold events. We present a strategy for deriving an adjustment to offset the gas estimate for days in an extreme cold event. By analyzing the forecast residuals for characteristic patterns and learning the statistics of the residuals, we build an adjustment model that predicts future values of residuals for approaching extreme events.

Chapter 4 considers the impact of prior day weather on daily gas demand. In this chapter, we explore the relationship between prior day temperature and current day gas demand. We derive an equation describing how prior day weather impacts demand. We show that by adjusting gas demand model with our prior day impact factor, the accuracy of the forecast can be improved especially during the unusually cold days.

# CHAPTER 2

## Identifying Extreme Cold Events Using Phase Space Reconstruction

Extreme cold events in natural gas demand are events characterized by unusual dynamics that makes modeling their behavior a challenging task. Natural gas utilities have to forecast well ahead the gas demand of their customers during extreme cold events to ensure adequate plans are in place to fulfil their customers' demand. To the natural gas utilities, extreme cold events represent high risk events given the associated huge demand. To improve the accuracy of gas demand forecast during extreme cold events, it is important to understand the nature of the unusual dynamics. In this chapter, we aim to identify extreme cold events in historical natural gas demand data. We present a semi-supervised pattern recognition algorithm that identifies extreme cold events in natural gas time series data. Using phase space reconstruction, the input space is mapped into a phase space. In the reconstructed phase space, events with similar dynamics are close together, while events with different dynamics are far apart. A cluster containing extreme cold events is identified by finding the nearest neighbors to an observed cold event. The learning algorithm was tested on natural gas consumption data obtained from natural gas local distribution companies. The identified events in each dataset are considered similar to the observed cold event.

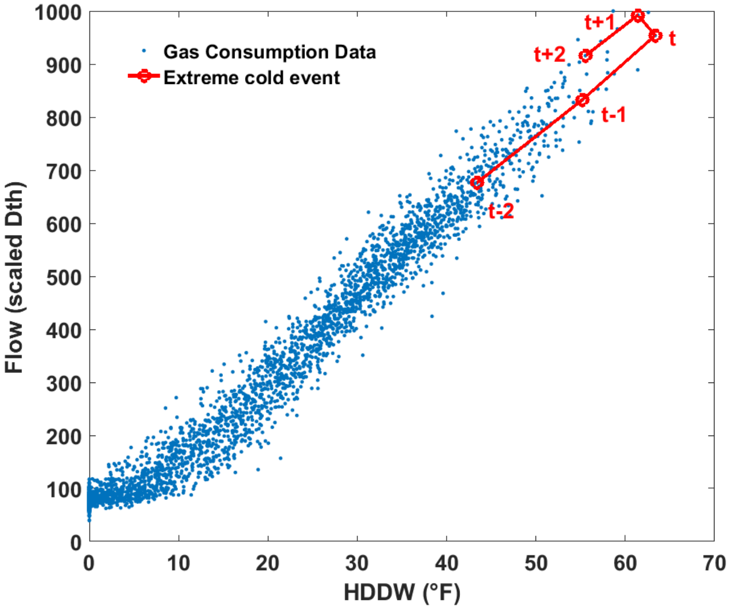
## 2.1 Motivation

The most important days in natural gas demand forecasting include the days when demand is at its peak. It is important to forecast gas demand accurately during this period because it helps in infrastructure, supply, and operational planning [37]. Gas demand increases as the temperature decreases, as most gas is used for space heating [66]. The highest gas demand occurs during extreme cold events. An extreme cold event is a multi-day event for which the temperature is below a given threshold (specified by 1-in-$n$ years) for several consecutive days with a characteristic hysteresis (see Figure 2.1a) in gas demand response. A 1-in-$n$ temperature denotes the temperature which occurs as infrequently as once every $n$ years. Extreme cold events represent one of the most challenging days of the year for operational gas forecasters because their gas delivery systems are operating near their maximum capacities. Considering the financial implications as well as physical limits to the amount of gas supply that can be made available during an extreme cold event, it is important to identify extreme cold events in natural gas demand data. Identifying these events enables us to conduct a more detailed study of extreme cold events, to understand the dynamics underplay during such events, and to develop better models to describe the response in gas demand during extreme events.

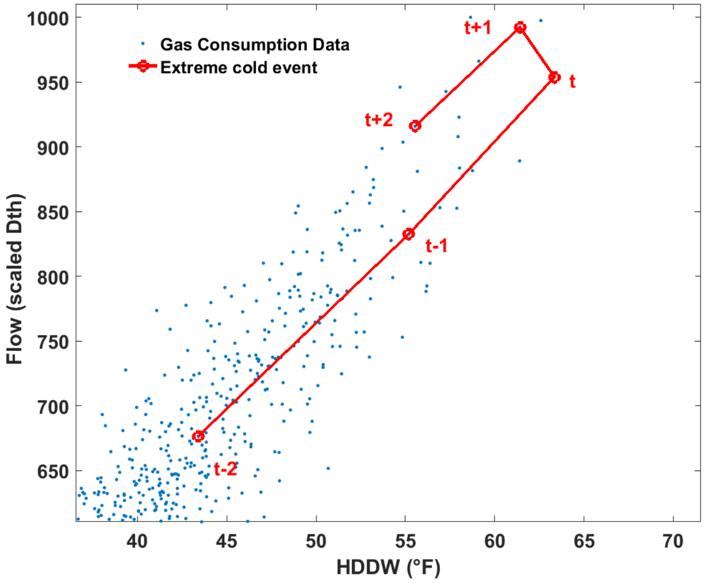
### 2.1.1 Behavioral Response

Extreme cold events are characterized by some interesting behaviors. Generally, gas demand varies almost linearly with temperature. For extreme cold events however, this relationship becomes non-linear. An unusual response in gas consumption in the form of hysteresis has been observed during the extreme cold events. Figure 1 shows the plot of daily natural gas consumption (flow) against wind-adjusted temperature (labeled $HDDW$), spanning a period of ten years. Figure 2.1b is a replica of Figure 2.1a with emphasis on the behavior of interest. The straight lines connect instances of natural gas consumption versus wind-adjusted temperature for five consecutive days. The days in the series identified by the lines represents the consumption for days $t-2$, $t-1$, $t$, $t+1$, and $t+2$, with $t$ being the coldest day in the event. The flow for the day after the coldest day $(t+1)$ is much higher than the flow for day $t$, even though the temperature is warmer. Apparently, people tend to use more gas even when it is not as cold as the day before.

Part of this response is due to thermodynamic effects, as heat transfer is a dynamic process. There is a certain time-lag relating the reported (outside) temperature to the actual temperature (inside the building). The lag factor depends on the building's insulation system. Murat [45] provides a good insight into the effect of thermodynamics on space heating in buildings. Attempts have been made to model the thermodynamics component by adjusting the forecast model for prior



(a)



(b)

Figure 2.1: An extreme cold event in natural gas consumption data for a certain region in the USA. The extreme event identified can be seen to exhibit a hysteresis effect as a result of unusual (human behavioral) response to extreme temperatures. The plot in (b) is an enlarged version of (a) with focus on the extreme event.

day weather effects as shown by [41] and [66]. The hypothesis here is that there is an unmodeled behavioral component, possibly due to human responses to extreme temperature and/or temperature changes [15, 16, 32], since the response to extreme cold events appears different from typical days.

### 2.1.2 Design Day Analysis

In addition to the behavioral response observed in gas demand, extreme cold events are associated with huge volume of gas demand due to the extreme low temperatures. Often, the highest gas demand is seen during extreme cold events. Due to the extreme weather conditions associated with extreme cold events, gas utilities are charged by the state commissions to ensure proper planning is put in place to meet customers' needs.

Design day analysis helps gas distribution utilities make adequate plans to meet the demands of their customers during extreme weather conditions. Part of the planning includes nominating the right amount of gas, and/or having enough gas in storage facilities. Also, when building gas distribution infrastructure, utilities need to know how large their distribution pipelines need to be to handle peak demands. The design of the distribution pipeline in a region is influenced by the highest probable demand for that region. This highest probable demand for which the system is designed is termed the design day demand.

In estimating the design and peak day demand accurately, it is important to understand the dynamics of demand during extreme cold events and compensate for the unusual behavioral response in the gas demand model.

### 2.1.3 Chapter Overview

In Section 2.1.1, we postulated that extreme cold events have different dynamics than usual days due to the unusual behavioral response. In identifying extreme cold events, we search for events in the data with similar dynamics to a known extreme event. The events are treated as temporal patterns.

We identify temporal patterns in natural gas data that correspond to extreme cold events. This is achieved by clustering the data based on dynamics. Natural gas demand data is high dimensional data, so that events with similar dynamics may not occupy the same cluster in the input data space. For effective clustering, a low-dimensional embedding of the data is performed using phase space reconstruction [52]. In the reconstructed phase space, events with similar dynamics are closer to each other, while those with different dynamics are far apart. Extreme cold events are identified by finding the events that are close to a known extreme cold event in the reconstructed phase space, using a nearest neighbor algorithm.

In the next section, we discuss important concepts on which the work presented in this paper is based, such as clustering and phase space reconstruction.

In Section 2.3, we describe our approach to identifying temporal patterns of extreme cold events in natural gas time series data. Pseudocode also is presented. In Section 2.4, we discuss the result of our pattern identification algorithm, and we present results obtained when the algorithm was evaluated on six gas demand data sets from different gas utilities.

## 2.2 Background on Method

This section provides a background on the various techniques used in this chapter. We will discuss concepts in clustering that are relevant to our specific task. We will also discuss phase space reconstruction - an approach in pattern recognition provide a few examples in literature where this approach has been used.

### 2.2.1 Clustering

Clustering algorithms are used in data mining and pattern recognition tasks where items are to be separated into groups. Items in the same group are considered similar, with similarity defined only in the sense of the particular application. Metrics used in determining similarity include distance (i.e., how close the points are), density (i.e., how compact points are), and connectivity. When using a distance function as a similarity metric, it is possible for similar points to be far apart in the input data-space, especially when dealing with high dimensional data.

In high dimensional spaces, distances between points are relatively uniform, so the concept of closeness is meaningless [59]. In clustering such high-dimensional data, it is customary to perform a low-dimensional embedding, mapping the input data space into a new space where closeness is properly defined.

### 2.2.2 Phase Space Embedding

One common technique employed in low-dimensional embedding of high dimensional data is phase space reconstruction based on Takens [60] time-delay embedding theorem. Takens' theorem gives the condition under which a dynamical system can be reconstructed from a sequence of observations of the state of the system. Sauer et al. [57] showed that for almost every time delay embedding with the appropriate selection of embedding parameters (dimension and time-lag), with a probability of 1, the reconstructed dynamics are topologically identical to the true dynamics of the underlying system. Hence, the underlying dynamics of a system can be captured fully in a reconstructed phase space (RPS).

This technique is able to reconstruct the underlying dynamics of any complex system and map it into a new lower dimensional space. Since the RPS is equivalent to the true dynamics of the system, points with similar dynamics are guaranteed to be close in this space, while less similar points are far apart [54, 56].

### 2.2.3 Temporal Pattern Identification Using RPS

The RPS approach was demonstrated by [53] to classify heart arrhythmia into one of four rhythms. An electrocardiogram signal was reconstructed in a phase space. The reconstructed phase was learned using a Gaussian Mixture Model (GMM) and classified using a Bayes classifier. Povinelli [53] showed that the RPS-based approach outperformed other frequency-based methods with an accuracy of up to 95%, compared to the 44% accuracy of the frequency-based method.

While most of the existing applications of the RPS approach deal with univariate time series where the temporal pattern to be identified appears in the same feature space, the RPS approach can be extended to multivariate time series. Zhang et al. [69] in detecting sludge bulking, a primary cause of failure in water treatment plants, used an RPS-based approach to identify multivariate temporal patterns characteristics of sludge bulking in sludge volume index (SVI) and dissolved oxygen (DO) time series. The SVI and DO time series data are embedded in a multivariate RPS. The embedding dimension and time-lag for each signal was estimated using global false nearest-neighbors and first minimum auto-mutual information [1]. A mixture of Gaussian models is used to cluster the multivariate reconstructed phase space into three distinct classes. The result of the RPS-GMM approach was compared to other methods and was shown to perform better than both ANN and Time Series Data Mining [49] approaches by at least 28%.

## 2.3 Identifying Extreme Cold Events

The techniques employed in identifying extreme events are similar to those described in Sections 2.2.1 through 2.2.3. This section discusses how the phase space reconstruction technique is applied to identify temporal patterns that correspond to extreme cold events in natural gas data.

Let an event be described as the dynamics between temperature and the corresponding natural gas demand over a series of five days. An extreme cold event is a five-day temperature-demand pattern whose dynamics are similar to that of a chosen target: coldest day in the historical dataset. An event is classified as an extreme cold event if the pattern associated with the unusual behavioral response described in Section 2.1.1 is detected. The natural gas dataset is a multivariate time series consisting of two separate time series; daily gas demand and daily temperature time series data. Let $s_t$ represent natural gas consumption for day $t$, and $HDDW_t$ be derived from the corresponding (wind-adjusted) temperature. An extreme cold event is a multivariate temporal pattern, defined as

$$p = \{s_1, s_2, ..., s_q; HDDW_1, HDDW_2, \ldots, HDDW_q\}\ , \tag{2.1}$$

with $p \in P \subseteq \Re^{2q}$, $q$ is the length of the temporal pattern. $P$ represents the pattern cluster. Given a multivariate time series $X = \{S(t); HDDW(t)\}, t = 1, 2, ..., n$, it is desired to identify all $p \in P$.

To identify all $p \in P$, $X$ is embedded in a multivariate reconstructed phase space in a way similar to [69]. The pattern cluster $P$ is identified using a nearest neighbor algorithm in the reconstructed phase space.

### 2.3.1 Data Preprocessing

The datasets used in this work were obtained from natural gas local distribution companies (LDC) across the USA. This data has been anonymized to protect the identity of the LDCs. Each dataset comprises ten years of actual gas consumption and weather data. The data is normalized prior to constructing a multivariate embedding. This ensures that $s_t$ and $HDDW_t$ are weighted equally in the reconstructed phase space such that the range of both $S$ and $HDDW$ is $[0, 1]$.

$$s_t = \frac{\max(s) - s_t}{\max(s)} \;, \tag{2.2}$$

$$HDDW_t = \frac{\max(HDDW) - HDDW_t}{\max(HDDW)} \;. \tag{2.3}$$

### 2.3.2 Multivariate Phase Space Embedding

The second step involves multivariate phase space embedding of the normalized time series data. According to [57], the appropriate selection of embedding parameters is necessary to ensure the reconstructed space is topologically equivalent to the original system. Takens' [60] original work argued that choosing embedding dimension $Q$ greater than $2m+1$, where $m$ is the dimension of the system's original state space, the time series can be completely unfolded in a phase space. Abarbanel [1] and Povinelli et al. [50] showed that useful information still can be extracted from the phase space by choosing a smaller $Q$. In most common applications [51, 53, 54, 69, 70], time-lag $\tau$ is estimated using the first minimum auto-mutual information, while dimension $Q$ is estimated using the global false nearest-neighbor technique. In [51], embedding parameters were selected based on the of length of the temporal pattern vector to be identified.

Our selection of embedding parameters is application-specific. The dimension $Q$ of the RPS and the time-lag $\tau$ at which to sample the signal are selected based on our domain knowledge. The selection of $\tau$ and $q$ is based on the length of the temporal pattern vector to be identified. We are interested in bitter cold events about five days long, so the inter-relationship between flow $S$ and wind-adjusted temperature $HDDW$ for five consecutive days interests us. Multivariate embedding is done by augmenting individual univariate RPS.

Flow time series $S(t)$ is embedded in a univariate RPS with time-lag $\tau = 1$ and dimension $Q = q = 5$. $S$ maps into $\Re^q$. The resulting phase space is a row vector

$$s = \left\{ s_1, s_2, \ldots, s_i, \ldots, s_{n-\tau(q-1)} \right\} \ , \tag{2.4}$$

$$s_i = \left\{ S_i \ S_{i+\tau} \ \ldots \ S_{i+\tau(q-1)} \right\} \ , \tag{2.5}$$

so that

$$s = \begin{bmatrix} S_1 & S_2 & S_3 & S_4 & S_5 \\ S_2 & S_3 & S_4 & S_5 & S_6 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ S_i & S_{i+\tau} & \ldots & & S_{i+\tau(q-1)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ S_{n-\tau(q-1)} & & \ldots & & S_n \end{bmatrix} .$$

$HDDW(t)$ is embedded in a univariate RPS with $\tau = 1$ and $Q = q = 5$ in a way similar to $S(t)$. The resulting phase space matrix

$$hddw = \left\{ hddw_1, hddw_2, \ldots, hddw_i, \ldots, hddw_{n-\tau(q-1)} \right\} \ , \tag{2.6}$$

$$HDDW_i = \left\{ HDDW_i \ HDDW_{i+\tau} \ \ldots \ HDDW_{i+\tau(q-1)} \right\} \ . \tag{2.7}$$

The univariate phase space matrices $s$ and $hddw$ have equal sizes. A multivariate RPS is formed by concatenating $s$ and $hddw$ such that the resulting multivariate phase space matrix is

$$\begin{bmatrix}
S_1 & S_2 & \dots & & S_5 & HDDW_1 & HDDW_2 & \dots & & HDDW_5 \\
S_2 & S_3 & \dots & & S_6 & HDDW_2 & HDDW_3 & \dots & & HDDW_6 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
S_i & & \dots & & S_{i+\tau(q-1)} & HDDW_i & & \dots & & HDDW_{i+\tau(q-1)} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
S_{n-\tau(q-1)} & & \dots & & S_n & HDDW_{n-\tau(q-1)} & & \dots & & HDDW_n
\end{bmatrix}$$

The overall embedding dimension $Q$ is the sum of the embedding dimensions of both variables, $Q = \sum_{i=1}^{2} q = 10$. Each row of the RPS matrix is a point in 10-dimensional space representing the dynamics of flow and temperature for five consecutive days.

Figure 2.2 shows a 3-dimensional projection of the 10-dimensional reconstructed phase space. Only three (namely $S(t-2)$, $HDDW(t-2)$, and $HDDW(t-3)$) of the 10 axes are shown for visualization purposes. Figure 2.2 also shows an event instance $e$ in the time series and its corresponding mapping in the RPS. The event $e$ shown in the time series plot has been reduced to a point in 10-dimensional space.

### 2.3.3 Nearest Neighbor Classifier

We desire to find the pattern cluster $P$ that corresponds to extreme cold events. This is achieved by classifying events into one of two classes: normal and extreme cold events. Classification is done in the reconstructed phase space obtained in

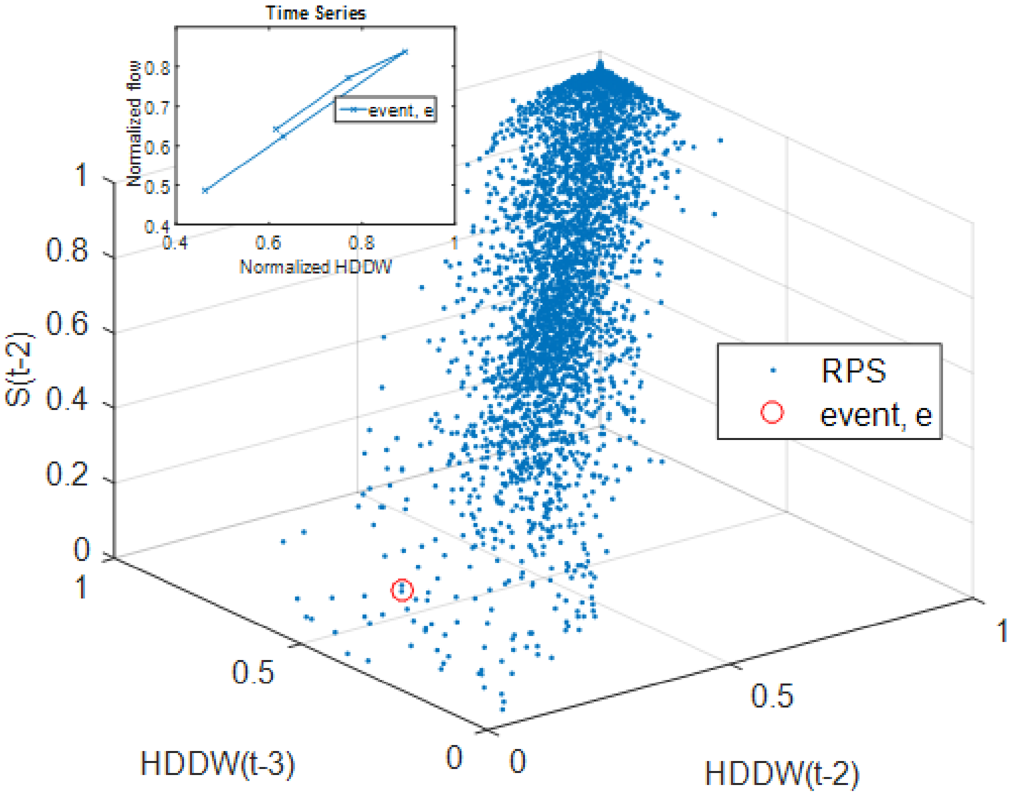


Figure 2.2: Reconstructed phase space built from natural gas consumption data. The overlayed plot (flow vs. *HDDW*) is an event instance $e$. In the reconstructed phase space, the event instance $e$ is represented by the red circle. The reconstructed phase space is a 10-dimensional phase space with axes $S(t), S(t-1), \ldots, S(t-4)$ and $HDDW(t), HDDW(t-1), \ldots, HDDW(t-4)$. The RPS plot shows only 3 of the 10 axes.

Section 2.3.2 using a nearest neighbor (NN) algorithm. This is possible because closeness can be defined in this new feature space.

Nearest neighbor is a nonparametric classification method based on the measurement of a point's similarity to a training set containing patterns for which class labels are supplied. A nearest neighbor classifier is an instance-based learning algorithm, i.e., it does not build a model through learning, but rather aggregates the values provided by the training patterns in the vicinity of the current point. A $K$-Nearest Neighbor ($k$-NN) classifier assigns a label to a point $x$ in the feature space based on the class assignment of its $k$-nearest neighbors. Decision is based on majority voting. This $k$-NN algorithm is supervised, requiring all training samples to have an assigned label.

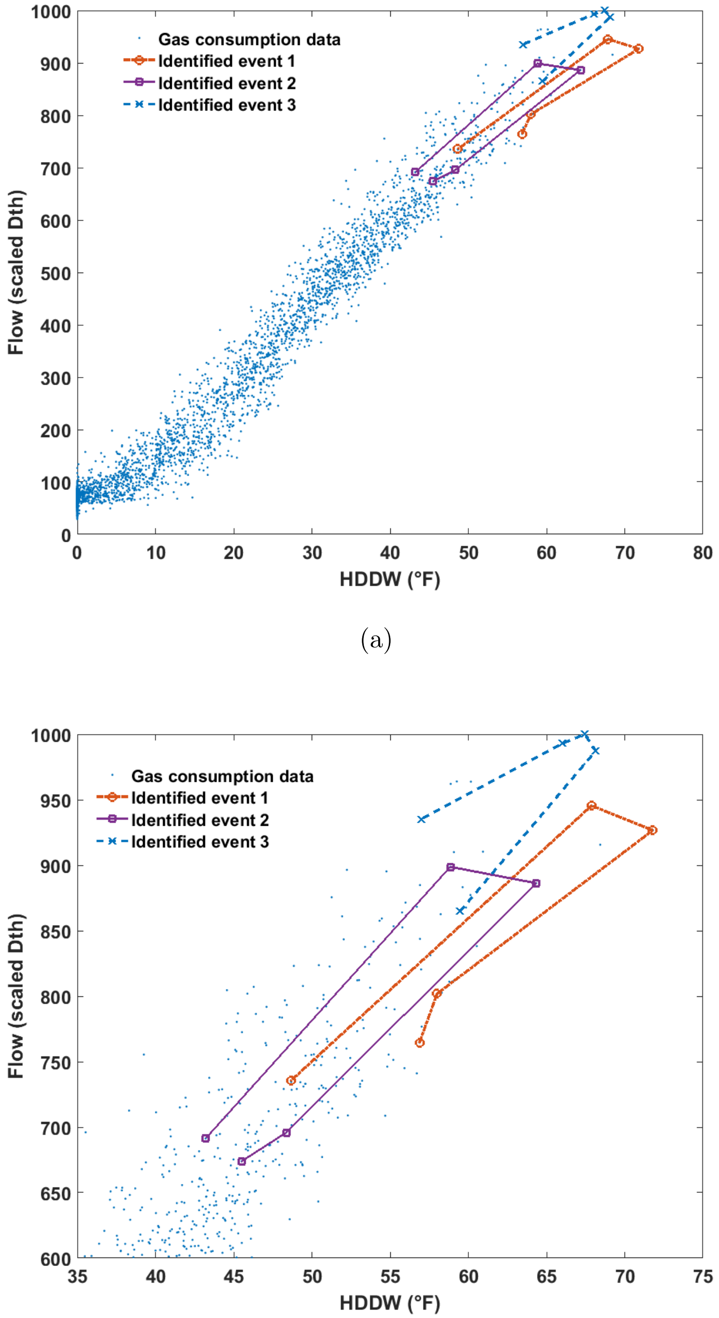
For an unsupervised task with unlabeled data, the $k$-NN algorithm no longer works. Identifying extreme events is an unsupervised task since there are no labeled datasets. To overcome this challenge, the unsupervised task is turned into a semi-supervised one by assigning a class label to one of the data points. This point is referred to as the pivot. The $k$-NN algorithm is modified to find the $k$ nearest neighbors to the pivot point (inclusive). The $k$ nearest neighbors discovered by this $k$-NN algorithm are assigned the same class label as the pivot. A known extreme cold event is chosen as the pivot, and the algorithm finds the $k$ closest events to the extreme cold event. Closeness of a point (to the pivot) is determined by computing its Euclidean distance $d(\text{pivot}, \text{event})$ from the pivot. The smaller the Euclidean distance, the higher the likelihood of the event being an extreme cold event and vice versa.

With the modified $k$-NN classifier described above, choosing the coldest event in the dataset as the pivot, the $k$-NN algorithm returned $k$ events that have the same dynamics as the observed coldest event. The coldest event is found by manually searching the reconstructed phase space for the event with the max $HDDW_{j+\frac{q-1}{2}}$ (i.e., lowest third day temperature for five-day events) and assigning it a class label: extreme event. Since the identification is done in the reconstructed phase space, the identified extreme events are mapped back to the original time series.

Figure 2.3 shows the flow and $HDDW$ time series with extreme events identified by the $k$-nearest neighbor classifier. In Figure 2.3, $k$ has been chosen as three for the purpose of presentation. Typical value of $k$ might be about two events per year of available data. The event identified by the circular marker is the pivot (coldest) event. The box and 'X' markers represent the other extreme events identified by the algorithm having a similar 'unusual response' to the pivot event.

### 2.3.4 Algorithms

The pseudocode of the RPS-$k$NN approach described in Sections 2.3.1 through 2.3.3 is provided in Algorithm 1. The **identifyExtremeColdEvents** function builds a multivariate RPS by merging two univariate RPS and calls the **classifyWithKNN** function to identify the extreme cold events. The **formUnivariateRPS** function builds individual RPS using the selected time lag $\tau$ and dimension $q$.



(a)

(b)

Figure 2.3: Three extreme cold events that have been identified using our RPS-$k$NN approach. The rightmost (coldest) event is chosen as the pivot. The two other events have been identified as the nearest neighbors to the coldest event in the reconstructed phase space. The plot in (b) is an enlarged version of (a) with focus on the extreme events.

**Algorithm 1** Reconstructed Phase Space - k Nearest Neighbor (RPS-kNN)

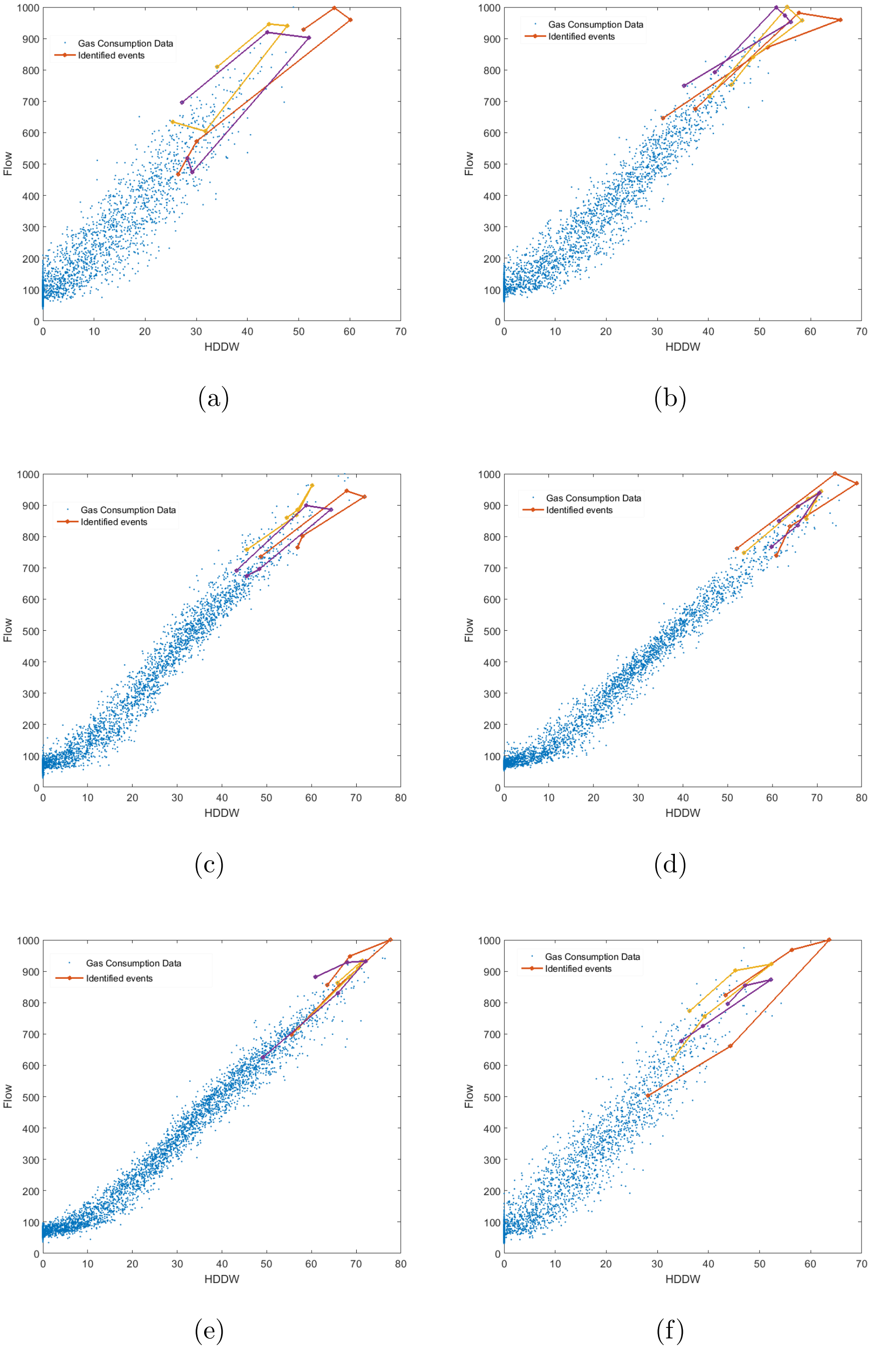
```
function identifyExtremeColdEvents(multivariateTimeseries, k)
    flow ← extract flow from multivariateTimeseries                ▷ Preprocessing
    HDDW ← extract wind-adjusted temperature from multivariateTimeseries
    normalizedFlow ← normalize flow
    normalizedHDDW ← normalize HDDW

    choose timelag $\tau$ and dimension $q$ based on domain knowledge          ▷ RPS
    rpsFlow ← formUnivariateRPS(normalizedFlow, $\tau$, $q$)
    rpsHDDW ← formUnivariateRPS(normalizedHDDW, $\tau$, $q$)
    rps ← merge rpsFlow and rpsHDDW to form a multivariate rps

    return extremeColdEvents ← classifyWithKNN(rps, k)   ▷ Classification
end function

function formUnivariateRPS(data, $\tau$, $q$)
    reconstructedPhaseSpace ← form a reconstructed phase space of data using the given $\tau$ and $q$ according to equations 2.6 & 2.7
    return reconstructedPhaseSpace
end function

function classifyWithKNN(rps, k)
    $x_i$ ← find coldest event and choose as pivot
    for all event $x_j$ in rps do:
        $d(i, j)$ ← compute the Euclidean distance
    end for
    d ← sort(d, asc)
    indexes ← return the indexes of the first k elements
    return extremeColdEvents ← re-map indexes in the phase space to time series
end function
```

## 2.4 Discussion

The RPS-$k$NN algorithm described in Algorithm 1 was tested on 20 datasets from different LDCs. Each dataset contains ten years of actual natural gas consumption data with the corresponding weather information. For each dataset, a multivariate reconstructed phase space is formed, and the coldest event in each dataset is chosen as the pivot. With the pivot chosen and $k$ fixed, the $k$-nearest neighbor classifier returns the $k$ most similar events to the coldest event. These events are considered to be the extreme cold events in the dataset. Figure 2.4 shows the events identified by the RPS-$k$NN algorithm for six datasets. Only three of the identified events are shown, for the sake of presentation.

### 2.4.1 Comparison with Previous Method

Previously, extreme cold events were identified by selecting the $k$ highest flow days in the historical data. The procedure involves searching through the daily gas demand data and picking the data point with the highest gas demand. Since extreme events are temporal data rather than a single data point, the days before and after the highest flow days were selected to form a temporal data. For instance, for a five day event, the highest flow day is selected as day 0, the two data points preceding day 0 are set to days $-2$ and $-1$, and the two days after the highest flow day are set to days 1 and 2. Once the first extreme cold event is selected, the search



(a) (b)

(c) (d)

(e) (f)

Figure 2.4: Events that have been identified as extreme cold events in natural gas consumption data using our RPS-$k$NN approach. For each plot, only 3 events are shown for the sake of presentation. Plots (a) through (f) show the identification result obtained when the RPS-$k$NN algorithm was executed on six datasets obtained from different natural gas local distribution companies in the United States. Each of the dataset used spans a period of ten years.

is repeated again to find the next extreme cold events. For each search iteration, previously selected days are excluded from the data. For example, for the first iteration the search space include all the historical data. The algorithm returns the indexes of the five days in the highest flow (extreme) event. For the next iteration (second highest flow events), the search space include all the historical data minus the five days in first identified event. For the third iteration, days identified in the first and second iterations are omitted from the data, and so on, until we get to the $k$th iteration.

This method of selection of extreme cold events does not consider any similarity in the dynamics of the identified events, unlike our RPS-$k$NN approach. Rather, the $k$ identified events are inspected visually by an expert who determines based on domain knowledge if those identified events are extreme cold events.

While both methods of identifying extreme cold events are fundamentally different, we still compare the events identified by both methods. Using both the RPS-$k$NN algorithm and the highest-flow method described in this section, setting $k$ to a constant value, we identify two sets of extreme cold events. We count the number of instances when both methods identify the same extreme cold events. This is done for values of $k$ between 10 and 30. The result of this comparison is shown in Table 2.1. The first row of the Table 2.1 shows, for $k = 10$, four out of the ten events returned by both methods are the same.

Table 2.1: Comparing the extreme cold events identified using the RPS-$k$NN algorithm with those identified using the highest-flow method

| Number of events identified | Number of same events | Number of different events | Ratio of same events | Ratio of different events |
|---|---|---|---|---|
| 10 | 4 | 6 | 0.40 | 0.6 |
| 12 | 5 | 7 | 0.42 | 0.58 |
| 15 | 7 | 8 | 0.47 | 0.53 |
| 20 | 9 | 11 | 0.45 | 0.55 |
| 30 | 10 | 15 | 0.40 | 0.60 |

### 2.4.2 Future Work

We have presented a pattern recognition technique based on phase space reconstruction and nearest neighbor algorithm to identify extreme cold events in natural gas data. Our RPS-$k$NN algorithm identifies extreme events by searching for temporal patterns in the data with dynamics similar to an observed extreme events (pivot event). Since the data is unlabeled, we initialize the algorithm by setting the pivot event as the coldest day in the historical data. This pattern identification approach is susceptible to initial condition since it ranks events based on their similarity to the pivot event that we select. Further study should be carried out for appropriate selection of the pivot event.

In this chapter, we have limited our consideration to five-day events. Other types of extreme events such as three-day and seven-day events should be considered as well. Our RPS-$K$NN algorithm is easily extensible to extreme events with different temporal length. While our algorithm was implemented for temporal pattern identification, with appropriate modification, it can be extended for classification task such as classifying extreme cold events into three, five, and seven-day events.

## 2.5 Conclusion

In this chapter, we discussed extreme cold events in natural gas demand data and how they are important to gas forecasters and gas distribution utilities. We pointed out unusual response that we have observed in gas demand data during extreme cold events, and we presented our RPS-$k$NN algorithm as a temporal pattern identification technique to identify extreme cold events in the gas demand data. Using the RPS-$k$NN algorithm, we identified extreme cold events in the data and compared the results to previous approach. The previous approach to identifying extreme cold events only considers the flow values while our RPS-$k$NN approach considers the flow-temperature dynamics in identifying extreme events.

In the next chapter, we will consider how the extreme cold events identified using our RPS-$k$NN algorithm might be used to improve the forecast accuracy of gas demand model during extreme cold events.

# CHAPTER 3

## Improving the Accuracy of Natural Gas Demand Forecasting By Analysis of Residuals

This chapter is presented as an extension of Chapter 2. In Chapter 2, we presented a semi-supervised pattern recognition algorithm to identify extreme cold events in natural gas data. In this chapter, we consider how the identified events can be used to improve the accuracy of gas demand forecast during extreme cold events.

Having identified temporal patterns corresponding to extreme cold events in gas data, we extend the work by analyzing the forecast residuals of the identified events for specific patterns. A residual learning model was developed to adjust the gas estimate for the unusual response observed during extreme cold events.
Figure 3.1 shows what this chapter hopes to achieve - a new estimate $\widehat{\widehat{s}}$ is derived by adding the estimate $\widehat{r}$ of residual $r$ to the initial estimated flow $\widehat{s}$.

### 3.1 Forecasting Extreme Events

In forecasting, extreme events are of special interest usually because they are at the tails of the historical distribution and difficult to model. To decision makers and operation managers, extreme events represent high risk events when they have to make mission-critical decisions. These extreme events are defined by domains and

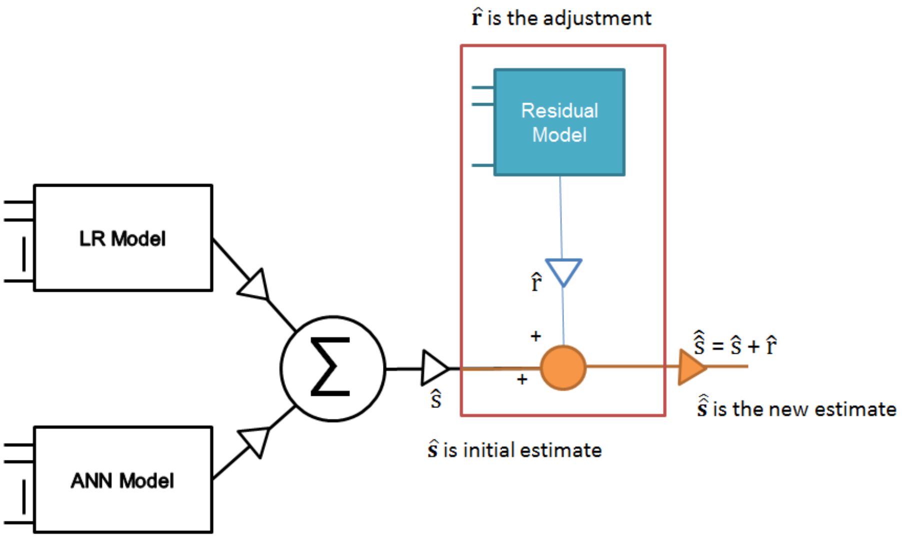

$\hat{r}$ is the adjustment
Residual Model
LR Model
ANN Model
$\Sigma$
$\hat{r}$
+
+
$\hat{\hat{s}} = \hat{s} + \hat{r}$
$\hat{s}$
$\hat{\hat{s}}$ is the new estimate
$\hat{s}$ is initial estimate

Figure 3.1: Adjustment model architecture for extreme cold events. Residual Model estimates the forecast residuals $\widehat{r}$ for days in an extreme cold event. A new estimate of gas demand is derived by adjusting the initial estimate $\widehat{s}$ with the residual estimate $\widehat{r}$.

are different for different domains. In financial forecasting, a trading edge that allows above normal returns would be termed an extreme event [49]. In flood forecasting and warning systems, a high impact flood that could cause damage to lives and property is an extreme event [8, 39]. For natural gas demand forecasting, periods of extreme cold weather represent extreme events.

### 3.1.1 Extreme Cold Events

An extreme cold event is a multi-day event for which the temperature is below a given threshold for several consecutive days with a characteristic hysteresis response (see Figure 3.2) in gas demand. These periods are associated with very high gas demand, making them very important to gas utilities [30]. Extreme cold weather events occur infrequently. They are usually specified by 1-in-$n$ years, i.e., the event is seen once in $n$ years. This means that extreme cold events are not sufficiently represented in historical data. Gas demand also has been reported to respond differently to weather during periods of extreme cold events [30]. The unusual response (shown in Figure 3.2) is in the form of hysteresis, resulting possibly from a behavioral response [28, 32].

This unusual response, combined with the rareness of the events, make it particularly challenging to forecast. In the next section, we give an overview of the model used to estimate daily gas demand and analyze the model's performance during extreme cold events.

### 3.1.2 Forecasting Gas Demand

In this section, we will describe a base model used to forecast daily gas demand. The base model is an ensemble of multiple linear regression (MLR) and artificial neural networks (ANN). The MLR model uses 13 input features which include $HDDW$, change in temperature from previous day ($\Delta HDD$), day of week, and autoregressive terms (lagged temperature and demand variables) [66]. The ANN model uses the same input features as the MLR model. The ensemble model was trained on 10 years historical data obtained from gas utilities in the USA. Testing was done on the remaining 10 years. The performance of the ensemble model was evaluated using mean absolute percentage error (MAPE). Because we are interested in the model's performance on extreme cold events, the model's estimate was plotted and overlaid on the actual demand. It is observed that this model has a consistent pattern in the error for days in the extreme cold event. For days before the coldest day in the event, the base model over-forecasts gas demand, and it under-forecasts days after the coldest day.

Figure 3.3 shows an extreme cold event. If $t$ is the index of the coldest day in the extreme cold event, on days $t$, $t-1$, and $t-2$, the dashed line (base estimate) is above the straight line (actual consumption), which means the demand forecast is more than the actual consumption for days before the coldest day. For days $t+1$ and $t+2$, the dashed line is below the straight line, which means that the gas demand forecast is less than the actual consumption for days after the coldest day. We will look more closely at the characteristics of the residuals in Section 3.2.2.

### 3.1.3 Quantifying Deviation

The work presented in this chapter offers a strategy for adjusting the base model estimate to improve the accuracy of the gas demand forecast. This strategy involves quantifying the deviation of the model (which is a result of unmodeled behavioral components) from the actual demand during extreme events. A computational

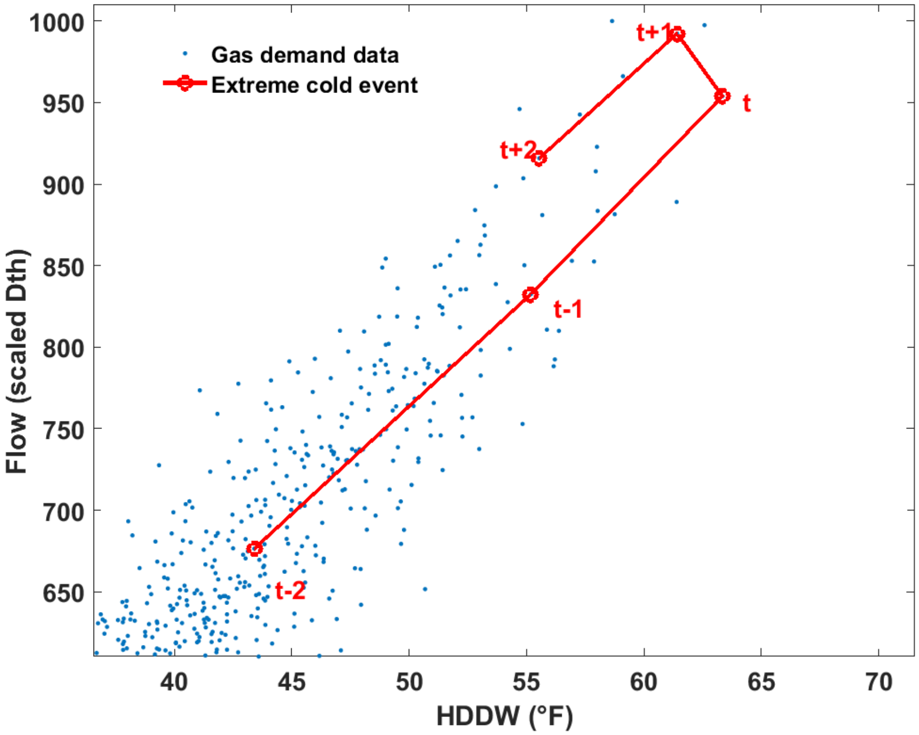


Figure 3.2: An extreme cold event in natural gas demand data. The extreme event identified can be seen to exhibit a hysteresis effect as a result of unusual response to extreme temperatures. It is expected that the highest flow is seen on day $t$ (being the coldest day). However, the next day (day $t+1$) has a much higher flow than the coldest day. Days $t+2$ and $t-1$ have approximately the same temperature, but their gas demands are almost 100 Dth apart.

model is built based on the statistics of this deviation to estimate the forecast residual during extreme cold events. This model is employed to estimate an adjustment to the underlying base model (see Figure 3.1).

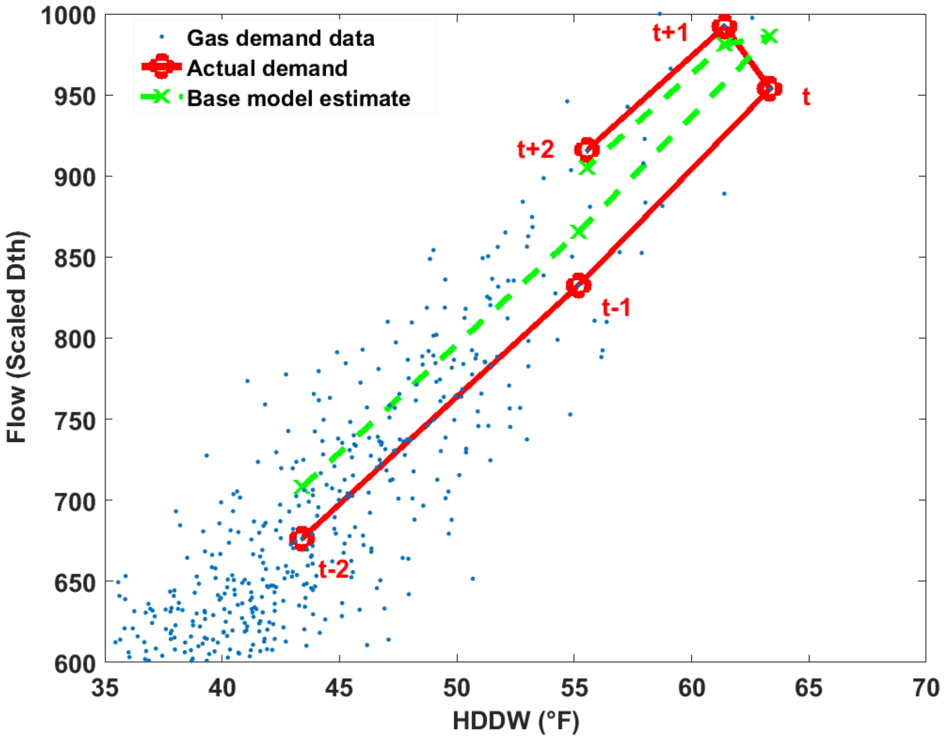


Figure 3.3: Performance of the base model during an extreme cold event. The base model (identified by the dashed line) over-forecasts gas demand for days before the coldest day $t$ and under-forecasts for days after.

### 3.1.4 Section Overview

Section 3.2 gives a theoretical basis for the methods used in this chapter.

Section 3.3 describes our adjustment model architecture to improve the estimate of gas forecast during extreme cold events. To estimate the adjustment during extreme cold events, the extreme cold events are first identified, as outlined in Section 3.3.1. Upon identifying the extreme events, the characteristics of the residuals (of identified events) are learned. Section 3.3.3 describes the two different residual models used in estimating the residuals. The model described in Section 3.3.3 to estimate residuals is used in Section 3.3.4 to forecast adjustments to the initial forecast $\widehat{s}$. Section 3.4 discusses the contribution of this work and offers recommendations to extend the work.

## 3.2 Background on Methods

This section provides a background on relevant time series and modeling concepts used in this work. We discuss nonlinear dynamic systems and the challenges they pose in modeling. We also explore previous work in which analysis of model residuals are used to improve the accuracy of time series forecasting. Lastly, we introduce the basic concepts of Partial Least Square (PLS) regression and discuss conditions under which PLS may be a better predictive model than the Ordinary Least Square method.

### 3.2.1 Nonlinear Dynamic Systems

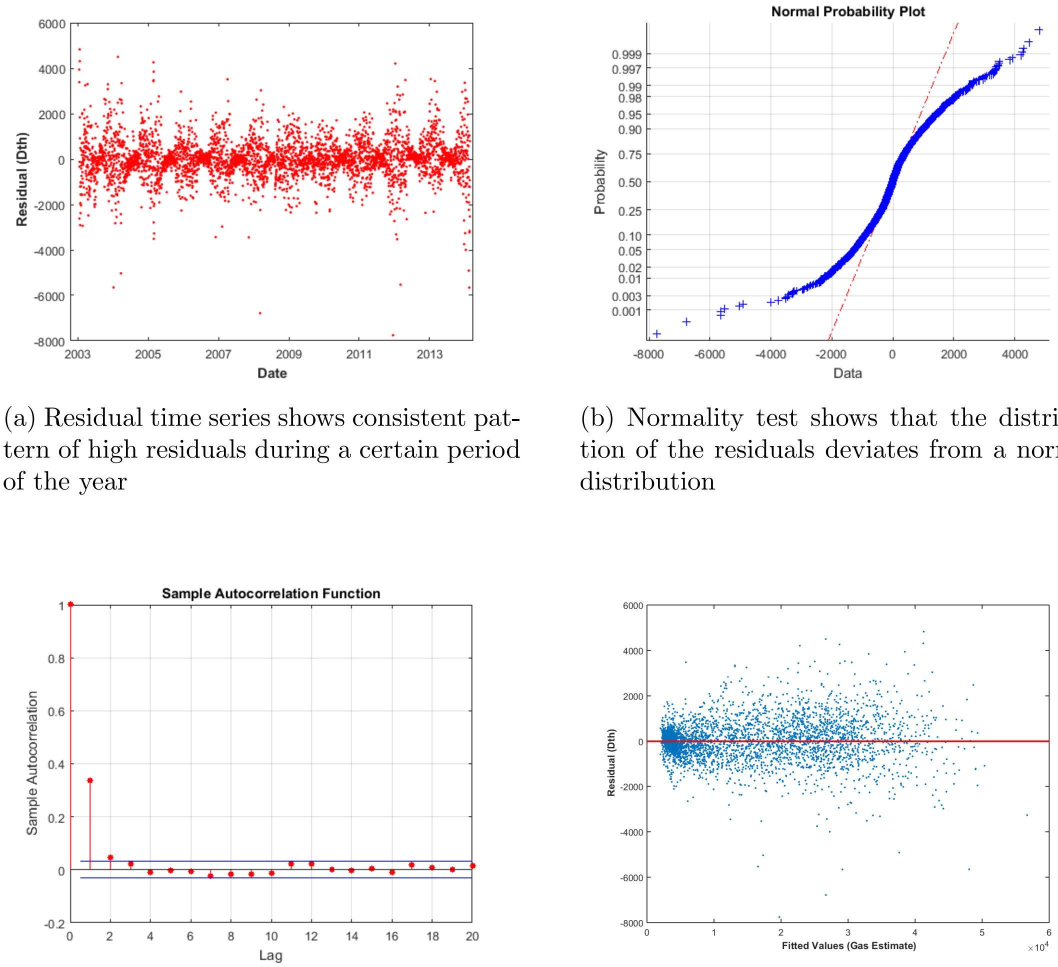
Nonlinear dynamic systems exhibit complex and seemingly unpredictable behavior [6, 10]. While these systems are deterministic, they often appear chaotic when observed in time series space. Forecasting such time series is a challenging task because chaotic systems are seemingly random in the time series domain. Many real systems exhibit this chaotic behavior such as a simple pendulum [36] or a person walking [10]. This form of complex behavior is seen also in natural gas demand data. Natural gas has been shown to have significantly different responses to different weather conditions [30].

### 3.2.2 Residual Analysis

In time series forecasting, a residual $r$ is the difference between the observed value of the dependent variable $y$ and its estimate $\widehat{y}$, i.e., $r = y - \widehat{y}$. Often, the residuals of forecasts are expected to follow a normal distribution with zero mean, and they often are assumed to have constant variance (homoscedasticity). If the residuals have a mean far from zero, then the model is said to be biased. Normally distributed residuals with zero mean should only be expected if the regression model has captured the input-output relationship of the underlying system sufficiently. In addition to being unbiased, a good regression model is expected to produce a residual that is uncorrelated, either with itself (auto correlation) or with other variables (cross correlation). If the residual shows correlation, bias, or patterns, then one or more predictors are not captured by the model.

Analyzing the residuals of the forecast obtained using the gas model described in Section 3.1.2, we obtain the plots shown in Figures 3.4a to 3.4d. The residual time series in Figure 3.4a shows a repeating pattern of high residuals during the heating season. The normality test in Figure 3.4b suggest the residuals do not follow a normal distribution. Figure 3.4c shows that the residual time series is correlated at a lag of one, even though the model contains an autoregressive component. Figure 3.4d shows the residual time series plotted against the estimate of our base model. We observe that the variance of the residuals varies with the values of the gas demand estimate. High values of the demand estimates are associated with large residual variances, while small demand estimates are associated with small residual variances. This behavior of the residuals violates the homoscedasticity assumption.

Statistical models developed to forecast time series data often overlook the characteristics of residuals [6]. Decisions regarding the selection of predictors, model order (in the case of mathematical modeling), and hyper-parameters (such as number of hidden neurons in a neural network) are based on 'educated guesses' and domain knowledge. It is impossible for any heuristic model to capture the complete characteristics of the time series data. It is often advised while validating models to inspect the residual plot for patterns. If the residuals of a time series prediction exhibit any discernible patterns, then crucial information can be learned from the residuals [6, 7]. Ardalani [6] demonstrated this by learning predictive information from the residual time series. In [6], future values of a chaotic time series were predicted by a neural network, and the model residuals were analyzed for chaotic behavior. The residuals from the initial prediction are treated as a new time series and learned using a time-delayed neural network. The network predicts the residual,



(a) Residual time series shows consistent pattern of high residuals during a certain period of the year

(b) Normality test shows that the distribution of the residuals deviates from a normal distribution

(c) Residual time series is autocorrelated at lag = 1

(d) Residuals suggest heteroscedasticity. The variance of the residuals varies by the values of the gas demand estimate

Figure 3.4: Testing the residuals of the gas forecast model. The residuals fails normality, correlation, and homoskedasticity assumptions, suggesting unmodeled components.

which is added to the initial prediction. This residual modeling approach was demonstrated to improve the accuracy of chaotic time series prediction by testing the approach on a Sunspot time series and two other publicly available chaotic time series [6].

The concept of learning to predict residual values to improve the accuracy of time series forecasting is only valid if the residual time series can be shown to follow a pattern. In the Sunspot example considered in [6], the residual was shown to be chaotic. To apply the residual learning approach to improve the accuracy of gas demand forecast, we have to demonstrate first that the forecast residuals have some pattern.

In Section 3.1.2, we introduced a base ensemble model to forecast daily gas demand. We also showed patterns (in the form of hysteresis) observed in the historical data during extreme cold events (see Figure 3.2). The rare nature of the extreme cold events makes it difficult for our base model to capture this pattern adequately, resulting in high forecast error during extreme cold events. Since the pattern is not captured by the model, the forecast residuals during extreme cold events are expected to reflect this pattern. Considering the noisy nature of the residual time series, coupled with the fact that the hysteresis pattern only occurs during extreme cold events, analysis of residuals in the context of residual learning is limited to days of extreme cold events.

Limiting the analysis of the residuals to extreme cold events poses a new challenge to modeling the residuals. Because extreme cold events occur infrequently, they are sparsely represented in the data. This means the amount of data on which to learn residuals is small.

### 3.2.3 Partial Least Squares

In predictive modeling, it often is desired to have a small feature-to-instance ratio. For instance, if the number of features is more than the number of observations in the training data, the model is likely memorize the data and is unable to generalize well on test data. For systems with large number of predictor variables (features), feature selection or dimensionality reduction becomes necessary [46]. Partial Least Squares (PLS) is a regression technique that combines dimensionality reduction and multiple linear regression, making it suited for modeling systems with large feature-to-instance ratios [38, 46, 63].

PLS works by constructing a smaller set of predictor variables (called components) and performs least squares regression on these components. The components are derived by finding linear combinations of input features that best explain the variance in the response variable [2, 46, 63].

To illustrate the performance of the Partial Least Squares regression on a dataset with small feature-to-instance ratio, we use the housing data set obtained from the University of California Irvine Machine Learning Repository [11, 55]. The data contains information about housing values in suburbs of Boston. The data has 506 instances, with 13 input features and one predictor variable (value of homes). The data was divided into training and testing data. To demonstrate the capability of PLS, we selected 20 data points randomly as training data to have small feature-to-instance ratio. The remaining 486 instances were used for testing.

The number of PLS components was determined using 5-fold cross-validation on the training data (20 data points). Based on the cross-validation error, the optimum number of component is six. Using this number of components, a new PLS model was built on the whole training data set. The learned PLS model is used to predict the value of homes in the test data.

The performance of the PLS model on test data was compared to MLR and ANN models. The MLR and ANN models were trained on the same 20 data points as the PLS model. Using the Mean Absolute Percentage Error as performance metric, the PLS model has a MAPE of 27.62%, MLR 39.41%, and ANN 44.83%. From the result obtained, it is clear that the PLS model has a better predictive ability than MLR and ANN for data with small feature-to-instance ratio.

### 3.3 Estimating Forecast Adjustment to Extreme Cold Events

We have shown in Sections 3.1.2 and 3.1.3 that extreme cold events exhibit behavior that is not accounted for in our base model, leading to the characteristic pattern seen in the residuals. In this section, we will discuss how residual analysis is used adjust the demand forecast to compensate for unmodeled components. The analysis is restricted to extreme cold events, since the pattern of hysteresis in the residuals is observed only during extreme cold events. Our goal is to derive an adjustment to offset the errors seen in the initial gas estimate on days in an extreme cold event and to show that the revised estimate improves the forecast accuracy during extreme cold events. We desire to derive an estimate $\widehat{r}$ of the residuals $r$ and add that value to the base model's flow estimate $\widehat{s}$ to arrive at a new estimate of gas demand $\widehat{\widehat{s}} = \widehat{s} + \widehat{r}$.

We assume the historical gas demand forecast $\widehat{s}$, produced by the base model is given, from which the residual $r = s - \widehat{s}$ is derived. The historical gas data used in this work, obtained from gas utilities across the US, spans data from 2003 to 2015. Data from 2003 to 2013 were used for all training and validation purposes, while the last two years of data were used for testing purposes.

Our task is to derive $\widehat{r}$ for days in an extreme cold event. Our algorithm (Adjustment Model) to achieve this is presented in the block diagram shown in

Figure 3.5. Since we are interested in learning residuals for extreme events only, the procedure starts with identifying the events of interest (Step 1), followed by estimating the residuals for days in the identified events (Steps 2 and 3), adjusting the base model forecast $\widehat{s}$ based on the residual estimate $\widehat{r}$ (Step 4), and finally fine-tuning the model (Step 5):

- Step 1: Identify extreme cold events
- Step 2: Calculate residuals of identified events
- Step 3: Learn residual of identified events
- Step 4: Adjust base model estimate
- Step 5: Optimize model parameters

Each of these steps will be discussed in the following subsections.

### 3.3.1 Identifying Extreme Cold Events

In Section 3.1.1, we introduced the nature of the extreme cold events observed in natural gas demand data and showed how extreme cold events exhibit dynamics different from usual days. To identify extreme cold events in the data, we find temporal patterns in the data with dynamics similar to an observed extreme cold

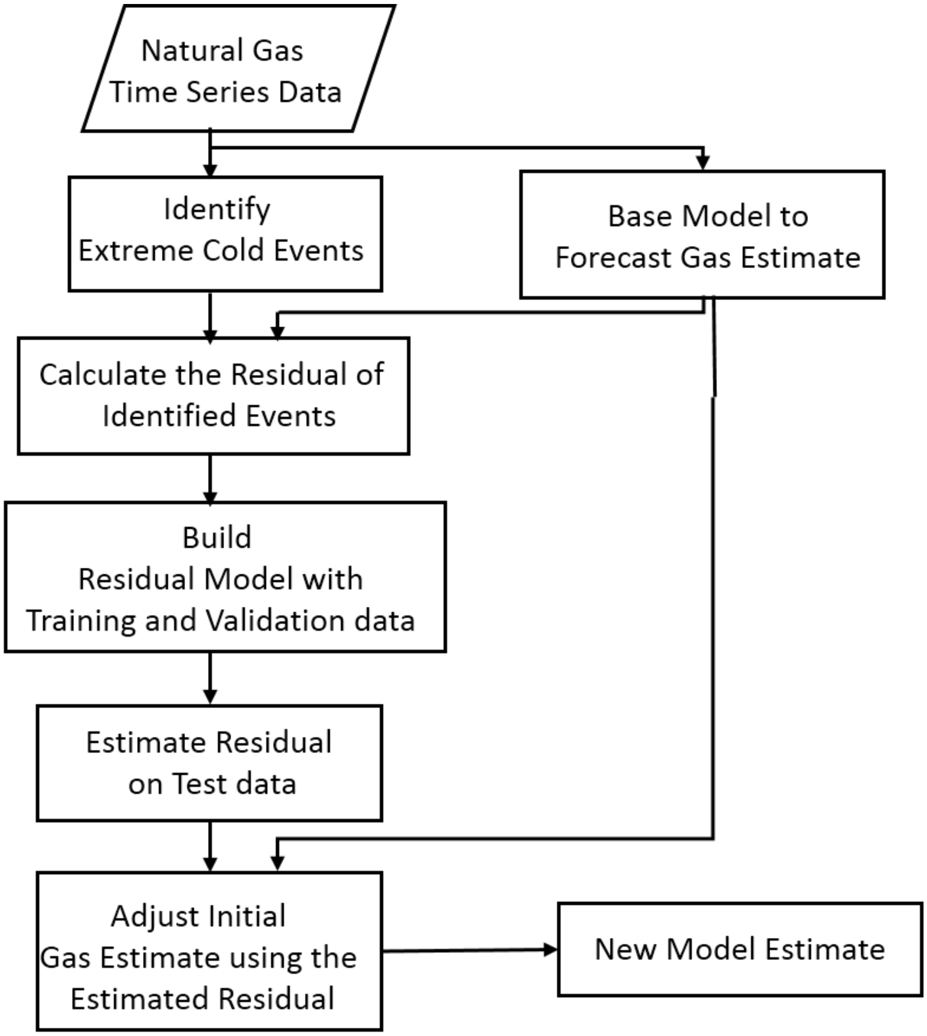


Figure 3.5: Block diagram for the Adjustment Model to improve the accuracy of natural gas demand forecast on extreme cold events. The base model produces the initial forecast. Extreme cold events are sampled from the data, and their corresponding residual are learned. An estimate of the residual is used as an adjustment to the initial base model estimate to derive the new demand estimate.

event. This was achieved using a combination of reconstructed phase space (RPS) and a $k$-nearest neighbour classifier ($k$NN) [30].

If we define an event by the interaction between temperature and the corresponding demand for five consecutive days, an event is classified as an extreme cold event if the pattern associated with the unusual behavioral response shown in Figure 3.2 is detected. The gas demand data is taken as a multivariate time series $X$ consisting of daily gas demand $s$ and daily (wind-adjusted) temperature $HDDW$. Let $s_t$ represent gas demand, and $HDDW_t$ be the corresponding (wind-adjusted) temperature for day $t$. An extreme cold event is a multivariate temporal pattern, defined as

$$p = \{s_1, s_2, ..., s_5; HDDW_1, HDDW_2, \ldots, HDDW_5\}\ , \tag{3.1}$$

with $p \in P \subseteq \Re^{10}$. P represents the pattern cluster. The identification problem is stated thus: Given a multivariate time series $X = \{S(t); HDDW(t)\}$, for $t = 1, 2, ..., n$, where $n$ is the is length of $X$, it is desired to identify all p $\in$ P.

To identify all p $\in$ P, X is embedded in a multivariate reconstructed phase space. A reconstructed phase space allows us to reconstruct the underlying dynamics of any time series data, given the appropriate selection of embedding parameters (dimension and time-lag) [57]. Since our goal is to cluster data based on dynamics, reconstructing the data in a phase space makes that possible. In the reconstructed phase space, events with similar dynamics are close to each other.

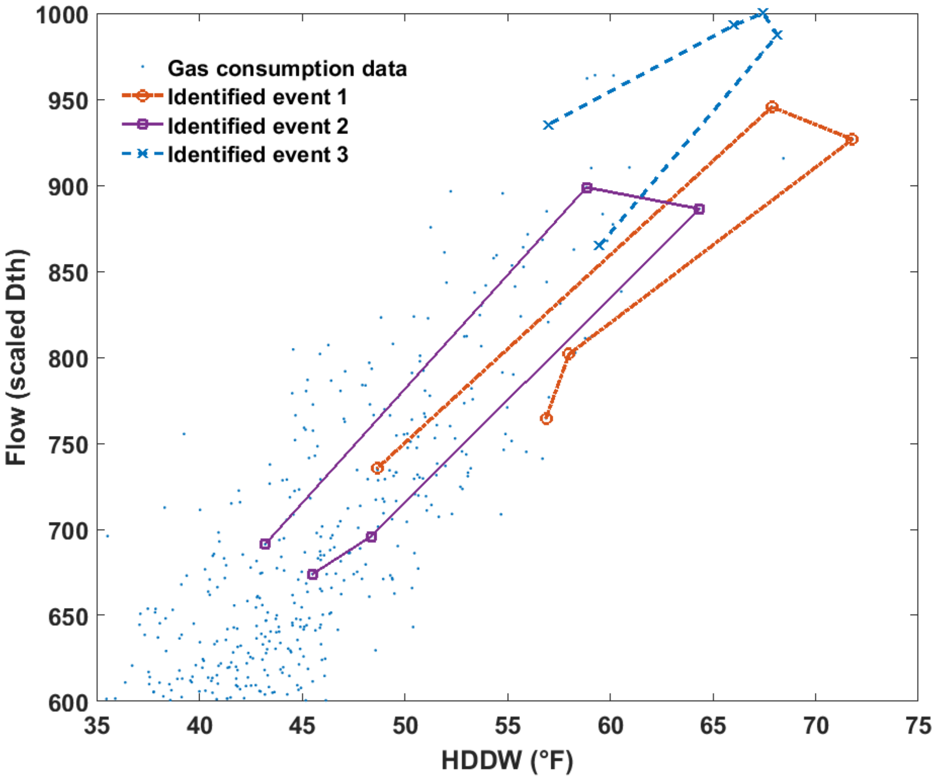


Figure 3.6: Three extreme cold events that have been identified using our RPS-$k$NN approach. The rightmost (coldest) event is chosen as the pivot. The two other events have been identified as the nearest neighbors to the coldest event in the reconstructed phase space.

Pattern cluster P is identified by finding $k$ nearest neighbors to an observed extreme cold (pivot) event in the reconstructed phase space.

Choosing embedding dimension of five and time lag of one, the gas demand data is reconstructed in a phase space. With $k$ set to 20, our RPS-$k$NN algorithm returned 20 extreme cold events. The complete details of the event identification step appears in [30]. The result of the identification step is shown in Figure 3.6. Only three events are shown in Figure 3.6 for the purpose of presentation.

### 3.3.2 Calculating Residuals of Identified Events

Having identified the extreme cold events in the data, the next step involves calculating the residuals of the forecast ($r = s - \widehat{s}$) for days in the identified events. The historical actual demands $s$ are known. The historical estimates of the demand $\widehat{s}$, derived from the base model are known also. We refer to $\widehat{s}$ as the base estimate.

Previously, an extreme event has been defined as a temporal pattern spanning five consecutive days. The forecast residuals $r$ on day $i$ for all days in the identified event are expressed as $r_i = s_i - \widehat{s}_i$ where $i = \{t-2,\ t-1,\ t,\ t+1,\ t+2\}$, such that $r_t$ is the forecast residual for day 3 in a 5-day extreme cold event, and usually, the coldest day in an event. Likewise, $r_{t+1}$ is the residual for a day after the coldest day, and $r_{t-1}$ is the forecast residual one day before the coldest day. The historical residuals $r_i$ are calculated for all days in the extreme events identified in Step 1 of the Adjustment Model in Figure 3.5.

The next step involves estimating the future values of residuals during extreme cold events. Analysis of the residual is done to understand the behavior of unmodeled dynamics. Understanding the characteristics of the residuals of the base model during extreme cold events is a precursor to deriving an adjustment to the base estimate.

### 3.3.3 Learning Residuals of Identified Events

Having obtained the residuals of the forecast for days in the identified extreme events, we build a computational model that learns the properties of the forecast residuals for the days in an extreme cold event. The learned residual model is used to estimate future values of residuals $\widehat{r}$ for days in an up-coming extreme event.

In predicting the residuals, two models are were developed. The first is a simple model that calculates the historical mean of the residuals and uses that mean as the expected value of the residual. The second model is a more flexible model that attempts to learn the relationship between residuals and selected features using Partial Least Squares (PLS) regression. The learned PLS model is used to predict future values of the residuals.

#### Expected Value Model

After calculating the residuals, the percentage residuals and average residuals by day are plotted in Figure 3.7. We observe that on days $t + 1$ and $t + 2$ (days after the coldest day), the mean values of the residuals are positive. For days $t - 2$ and $t - 1$ (days before coldest day), the residuals are negative. The expected value approach uses this mean value as the estimate of the residuals.

For each day, a distribution of the residuals for all identified events is built.

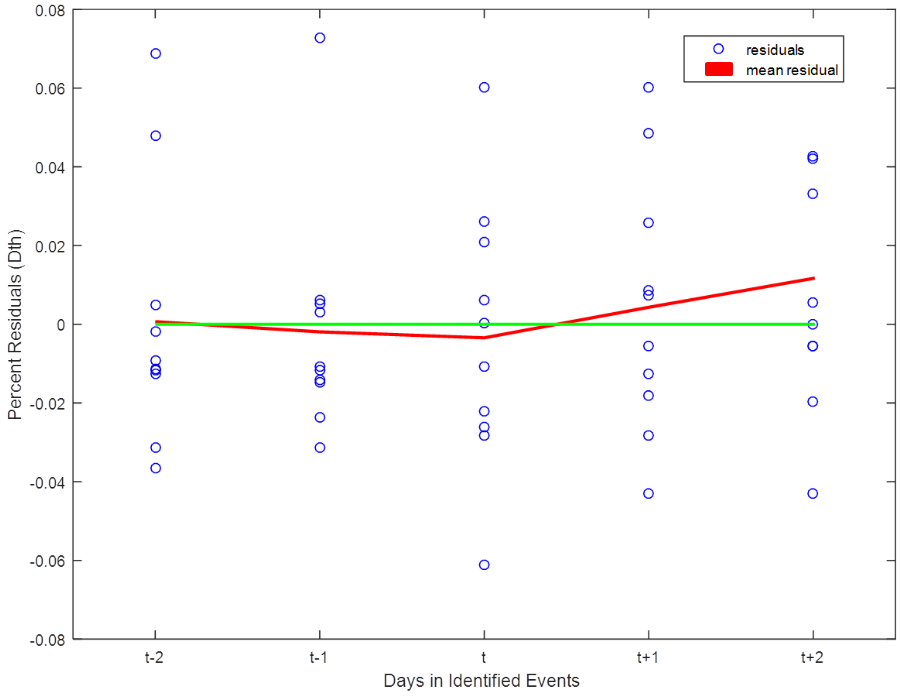


Figure 3.7: Percentage residuals by day for ten identified extreme cold events.

Figure 3.8 shows the histograms for all five days in the events. For day $t$, We estimate the residual $\widehat{r}_i$ using the expected value $E[r_i]$ of each distribution.

$$r_i = s_i - \widehat{s}_i \ , \tag{3.2}$$

$$\widehat{r}_i = E[r_i] \ . \tag{3.3}$$

### Partial Least Squares Model

The Expected Value model uses only the statistics of the residuals to estimate future values of residuals $r$. In this section, we attempt to build a residual model

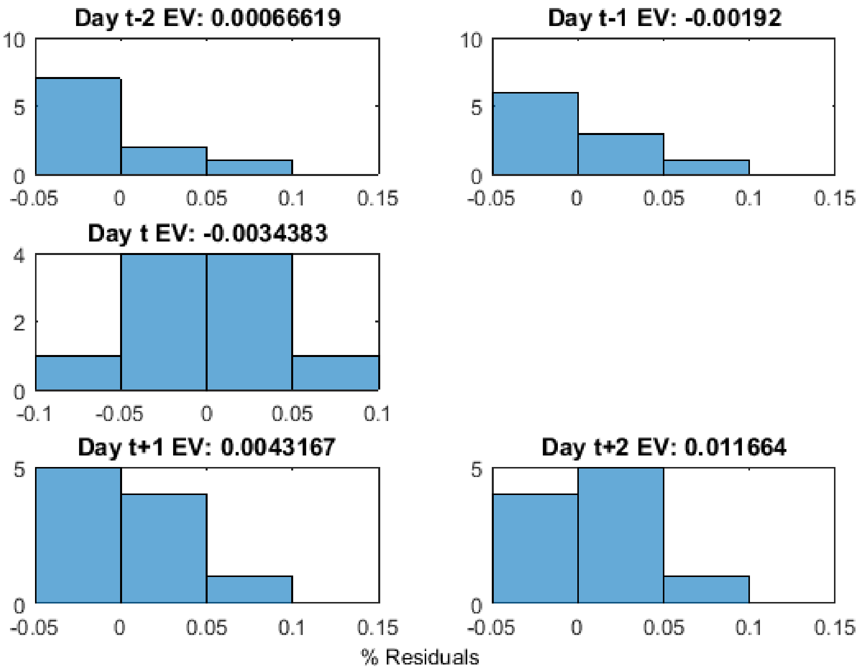


Figure 3.8: Histograms of the residuals $r_i$ for days in the identified extreme cold events. The Expected Value for each day is estimated from the mean of the distribution. Day $t$ represents the third day in a 5-day event (often the coldest day). Days $t+1$ and $t+2$ represent days after the coldest day, while days $t-1$ and $t-2$ represent days before the coldest day.

that is feature dependent. However, we have two major challenges. We are limited by the number of samples available, i.e., the number of extreme cold events identified. Also, it is not evident what predictor variables to use.

To tackle the first challenge, we use a Partial Least Squares model to learn the characteristics of the residual. We showed in Section 3.2.3 that a PLS model is able to learn on a small sample size. In the event identification step, we set $k = 20$, so that our sample size equals 20 cold events. Points in the reconstructed phase space derived in Section 3.3.1 are used as input features to the PLS model. Points in the RPS were chosen as input features since they are a representation of the state of the system. For each identified event, we have ten input features, which are the components of the event in the 10-dimensional reconstructed phase space. The residual $r$ is the response variable. The PLS residual model is learned on the identified events. The learned PLS model is used to estimate future values of the residuals $\widehat{r}$. The Partial Least Squares regression tool (**plsregress**) in **MATLAB** is used for the analysis in this section.

The PLS model creates a new set of predictors (components) which are linear combinations of our input features (i.e., points in the RPS). The number of components was determined through cross validation. From the cross validation plot in Figure 3.9, as the number of components is increased, the Mean Square Error (MSE) increases initially up until three components. This means using a PLS model with fewer than four components would not help. A PLS component of zero seems to suggest that our adjustment model will not result in any improvement. The PLS model with seven components ($PLS_7$) gave the least Mean Square Error (MSE) on validation. The $PLS_7$ model, learned on the residuals of the identified events is used to estimate future values of residuals $\widehat{r}$.

To estimate the residuals $\widehat{r}$ for each day $i$ in an extreme event, five PLS models were built using the procedure described in this section, one for each day of

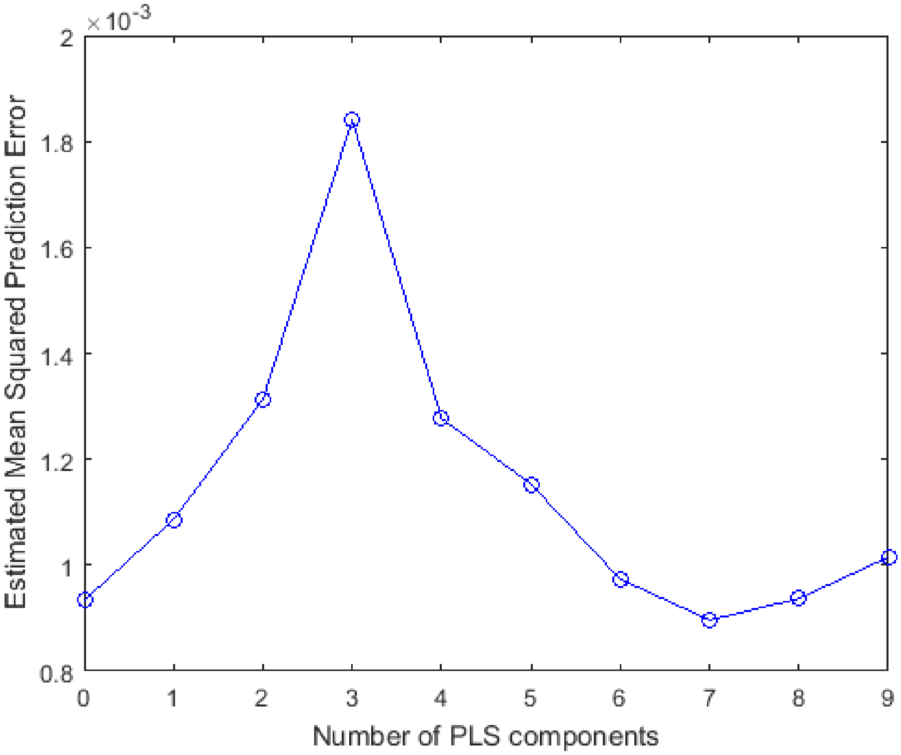


Figure 3.9: Partial Least Squares model cross validation error. The PLS Model with seven component gave the least MSE.

an event. All PLS models were trained on the same input features, but with different residuals. The PLS model for day $i$ was trained on $r_i$ and used to estimate $\widehat{r_i}$.

### 3.3.4 Adjusting the Base Model Estimate

From Section 3.3.3, we derived the estimates of the residuals $\widehat{r_i}$ for all days in an event. In this section, $\widehat{r_i}$ is used as adjustment to the initial base estimates $\widehat{s_i}$ to derive new estimates of gas demand, $\widehat{\widehat{s_i}} = \widehat{s_i} + \widehat{r_i}$. The adjustment is applied only on days in an approaching extreme cold event.

In Section 3.3.3, we estimated residuals using Expected Value (EV) and Partial Least Squares models. The residuals $\widehat{r_i}$ produced by both models were used separately in deriving $\widehat{\widehat{s_i}}$. The impact of our adjustment was evaluated by comparing $\widehat{\widehat{s_i}}$ to the actual gas demand $s_i$. The Mean Absolute Percentage Errors (MAPE) between $s_i$ and $\widehat{\widehat{s_i}}$ for all days $i$ were evaluated.

### 3.3.5 Optimizing k and number of PLS components

So far, the learning of residuals and performance evaluation has been based on the events identified in Step 1. The effectiveness of our RPS-$k$NN algorithm for identifying extreme cold events is dependent on the value of $k$. The RPS-$k$NN algorithm returns $k$ closest events to an observed extreme cold event. As $k$ increases, the likelihood of the identified events having similar dynamics to the observed extreme event decreases [30]. On the other hand, we want $k$ to be large enough to ensure we have data, so that our adjustments generalize well. The optimum value of $k$ is chosen such that our adjustment model, built on $k$ identified events, leads to improved estimates of gas demand during extreme cold events. Thus, our optimization problem (expressed in Equation (3.4)) becomes finding $k$ that minimizes the validation MAPE between the $s_i$ and $\widehat{\widehat{s_i}}$ for day $i$.

$$\min_{k} \left( \frac{1}{k} \sum_{j=1}^{k} \frac{s_{ij} - \widehat{\widehat{s_{ij}}}}{s_{ij}} \right) . \tag{3.4}$$

For the PLS-adjustment model, the optimization considers the number of PLS components (also called latent vectors LV) as well. Let nLV represent the number of PLS components, the optimization is expressed as,

$$\min_{k,\ \text{nLV}} \left( \frac{1}{k} \sum_{j=1}^{k} \frac{s_{ij} - \widehat{s_{ij}}}{s_{ij}} \right) . \tag{3.5}$$

We set $k$ ranging from 5 to an arbitrarily large value. For the PLS model, the number of PLS components is set to values ranging from 1 to $\min(9, k-1)$. Using Equation (3.4) for the EV model and (3.5) for the PLS model, we repeat Steps 1 to 4 while varying $k$ and $nLV$. An exhaustive search is performed to find the best adjustment model. The adjustment model that results in the least MAPE on validation data is chosen. This is done for all five days in an event, yielding five adjustment models, one for each day in an event. Figure 3.10 shows cross validation error for day 3. In Figure 3.10, the optimum $k = 11$ while the optimum $nLV = 4$, selected using Equation (3.5). This combination pair of $k$ and $nLV$ gave the minimum validation MAPE for $k - nLV$ space searched.

### 3.3.6 Estimating Adjustment for Future Events

From the last section, we trained and validated forecast adjustment models for extreme cold events. In this section, the validated models will be used to predict

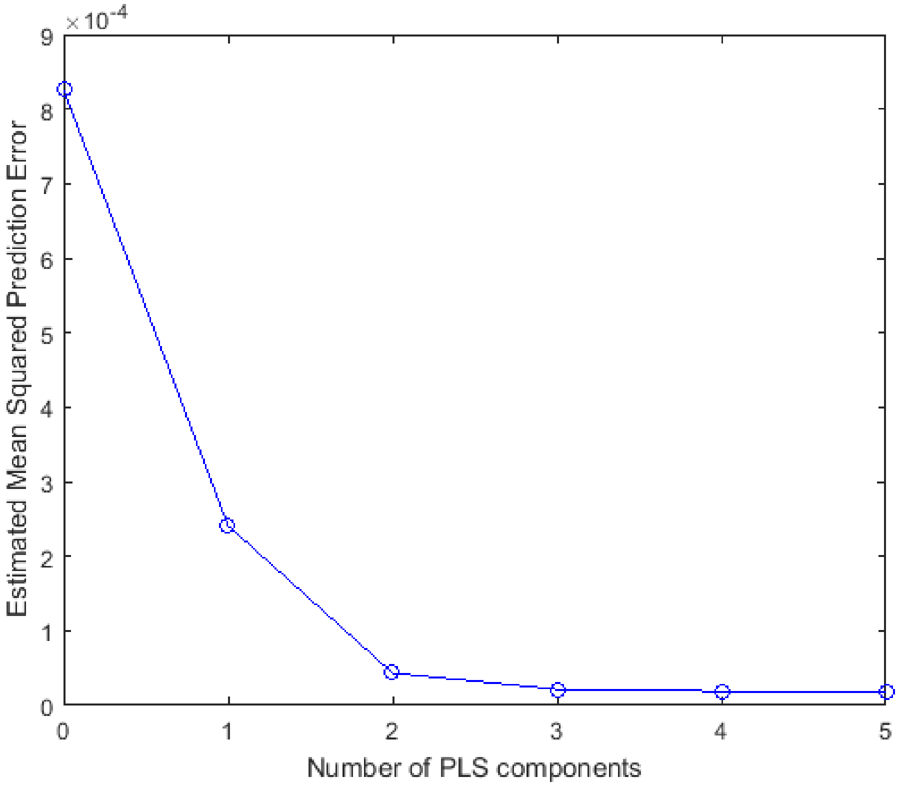


Figure 3.10: Partial Least Square cross validation error for optimum $k = 11$.

forecast residuals for upcoming extreme events. The analysis in this section will be done only on test data.

The first step in estimating adjustment to an upcoming extreme cold event is to identify if an upcoming event is classified as extreme cold event. If an upcoming event is an extreme cold event, then we apply our adjustment model to offset the base model's gas estimate; otherwise, we do nothing.

In identifying extreme cold events in the historical data in Section 3.3.1, we performed an RPS embedding on the historical data and used a nearest neighbor

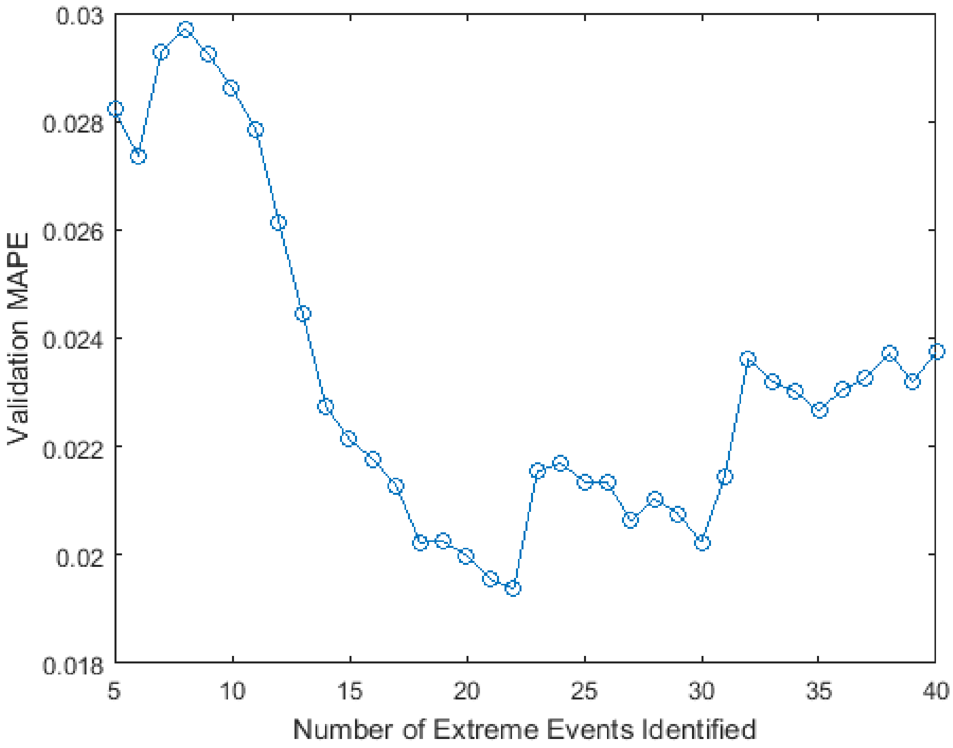


Figure 3.11: Expected Value validation error. Until $k = 22$, the MAPE decreases as the number of events increases. For values of $k$ above 22, the validation MAPE continues to increase.

algorithm to identify $k$ events that are closest to an observed extreme cold (pivot) event in the RPS. The nearest neighbor classifier used the Euclidean distance function as similarity measure.

To classify an upcoming event, we perform an RPS embedding of the event, calculate its Euclidean distance relative to the pivot event (used in Section 3.3.1) in the reconstructed space. If the Euclidean distance is within the Euclidean distance of the $k$th event, we classify the event as an extreme cold event.

Our test data consist of actual gas demand $s$, gas estimate from our base model $\widehat{s}$, and wind-adjusted temperature $HDDW$ from $2013 - 2015$. Upon embedding this data in a reconstructed phase space according to the description in Section 3.3.1, we compute the Euclidean distance from the pivot event for all points in the RPS. Events whose Euclidean distance fall within the radius of the $k$th event centered at the pivot are classified as extreme cold events. For each of the events identified, we use the validated adjustment models to estimate the expected forecast residuals $\widehat{r}$ for all days in the event. Since five adjustment models were trained for each of the five days in an event, the number of events $(k)$ on which the adjustment models were trained is expected to vary. Although our test data set is fixed (2013-2015), the number of events identified in the test data also varies for each of the adjustment models since the $k$ varies.

The estimated residuals $\widehat{r}$ are added to the initial base model's gas demand estimate $\widehat{s}$ to arrive at a new gas demand estimate $\widehat{\widehat{s}}$. The performance of the adjustment model was evaluated by comparing $\widehat{\widehat{s}}$ to $s$ using the Mean Absolute Percentage Error. The results are presented in Tables 3.1 and 3.2. The MAPE values shown in Tables 3.1 and 3.2 are estimated on the extreme cold events identified in the test data. Since the identification process is dependent on the $k$, the number of test events identified changes with $k$ for each adjustment model. The base model MAPE is calculated on the same test event as the adjustment model.

From the results obtained, we observe that our adjustment model did not

Table 3.1: Comparing MAPE of the adjustment models with the MAPE of the base model for all five days in the identified events

| | Base Model | PLS | PLS % Improvement | k |
|---|---|---|---|---|
| Day 1 | 0.0367 | 0.0382 | -4.09 | 30 |
| Day 2 | 0.0230 | 0.0308 | -33.91 | 19 |
| Day 3 | 0.0297 | 0.0258 | +13.13 | 21 |
| Day 4 | 0.0363 | 0.0290 | +20.11 | 12 |
| Day 5 | 0.0369 | 0.0406 | -10.03 | 30 |

Table 3.2: Comparing MAPE of the adjustment models with the MAPE of the base model for all five days in the identified events

| | Base Model | Expected Value | EV % Improvement | k |
|---|---|---|---|---|
| Day 1 | 0.0380 | 0.0407 | -7.11 | 40 |
| Day 2 | 0.0221 | 0.0250 | -13.12 | 5 |
| Day 3 | 0.0297 | 0.0300 | -1.01 | 22 |
| Day 4 | 0.0352 | 0.0339 | +3.69 | 7 |
| Day 5 | 0.0370 | 0.0368 | +0.54 | 25 |

improve upon the initial gas estimate for all the days in the event. Sometimes it shows improvement; other times, it increases the forecast error. The PLS model gave the highest improvement on day three with 13.1% reduction in MAPE, and the worst performance on day two with 33.9% increase in MAPE on the test data. For the EV model, the best performance is seen on day four with 3.7% reduction in MAPE, while the worst performance is seen on day two with 13% increase in MAPE. Overall, we conclude that our model adjustment technique is unable to learn effectively the forecast residuals during extreme cold events.

## 3.4 Conclusion

In this chapter, we presented a residual learning technique by which the accuracy of gas demand forecast can be improved during extreme cold events. We started by establishing a basis for residual analysis. We pointed out a pattern of residuals observed during extreme cold events. We highlighted two major challenges that make it difficult to forecast extreme cold events accurately. One, extreme events are rare and under-represented in the historical data. Two, we showed that extreme events have dynamics different from usual days, possibly due to human behavioral response to extreme low temperature. While we acknowledge that it is difficult to model gas demand accurately during extreme events due to the two reasons mentioned above, extreme events pose a major business risk to gas utilities. Hence, the motivation for finding a way of improving the forecast accuracy during extreme cold events.

To improve the accuracy of gas demand forecast during extreme cold events, we devised a model adjustment architecture that learns the forecast residuals during extreme cold events. The learned residual model is used to predict future values of residuals which are used to offset the initial base model's estimate of gas demand. Adjustment models were built for each day in an event on training and validation data. The adjustment models were tested on extreme cold events identified in the test data, and their performance was evaluated using the Mean Absolute Percentage

Error. The result presented in Tables 3.10 and 3.11 showed no conclusive evidence that our adjustment model architecture will improve the forecast accuracy of future extreme cold events.

### 3.4.1 Another forecast improvement technique

In this chapter, we attempted to learn to improve the accuracy of gas estimate during extreme cold events using a residual learning technique. The result we obtained was not consistent. In some instances, we obtained increased accuracy; other instances, the forecast accuracy decreased. In Chapter 4, we will consider another technique by which the accuracy of gas demand forecast can be improved. This techniques works by deriving a good approximation of the complex relationship between temperature and gas demand and uses that information to develop a better gas estimate model. This technique, discussed in Chapter 4, resulted in consistent improvement in forecast accuracy not only on extreme cold event but also several unusual day types.

# CHAPTER 4

# Impact of Prior Day Weather Sensitivity on Natural Gas Demand

This chapter presents yet another method by which the accuracy of gas demand forecasts can be improved. In this chapter, we will consider the impact of prior day weather on daily gas demand. Our analysis involves deriving a function that describes this impact factor and building a gas demand model that uses the information learned from our analysis to improve the accuracy of gas demand forecasts.

## 4.1 Motivation

The demand for natural gas demand is driven mostly by temperature. Gas demand can be divided into two components: baseload and heatload. Baseload demand is the minimum quantity of natural gas that the gas distribution companies (utilities) must make available to their customers, independent of temperature. The baseload is fairly constant, and its uses include cooking, drying, and heating water. Heatload is the amount of gas needed for space heating during periods of cold weather. The heatload varies approximately linearly with temperature. As the temperature decreases, people burn more gas to heat their buildings. Figure 4.1 shows historical

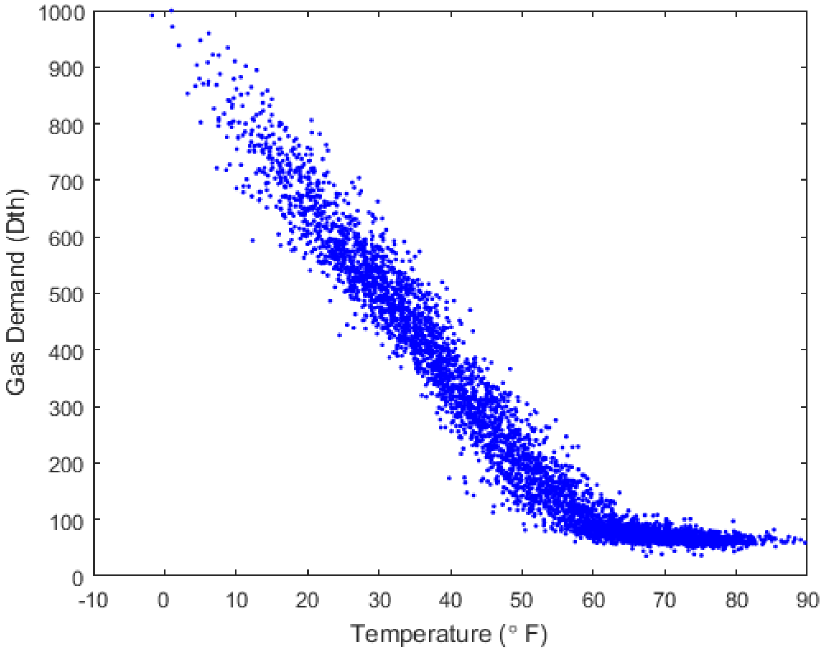


Figure 4.1: Historical daily gas demand against daily temperature. Gas demand is temperature independent for temperatures above 65°F. A Large percentage of gas use is driven by temperature.

demand of gas plotted against temperature for a typical operating area. Figure 4.1 shows that temperature is highly correlated with gas demand for temperatures below about $60 - 65$°F.

Statistical models used in forecasting daily gas demand use daily temperature as one of the predictor variables. Often, the temperature variable is transformed to Heating Degree Day $HDD = \max(0, T_{ref} - T)$. In our case, $T_{ref}$ is chosen as 65°F so that for temperatures above 65°F, $HDD = 0$. Figure 4.2 shows that gas demand is nearly linear with $HDD$. Other predictor variables often used include wind speed, cloud cover, day of the week, holidays, and trends due to technological, economic,

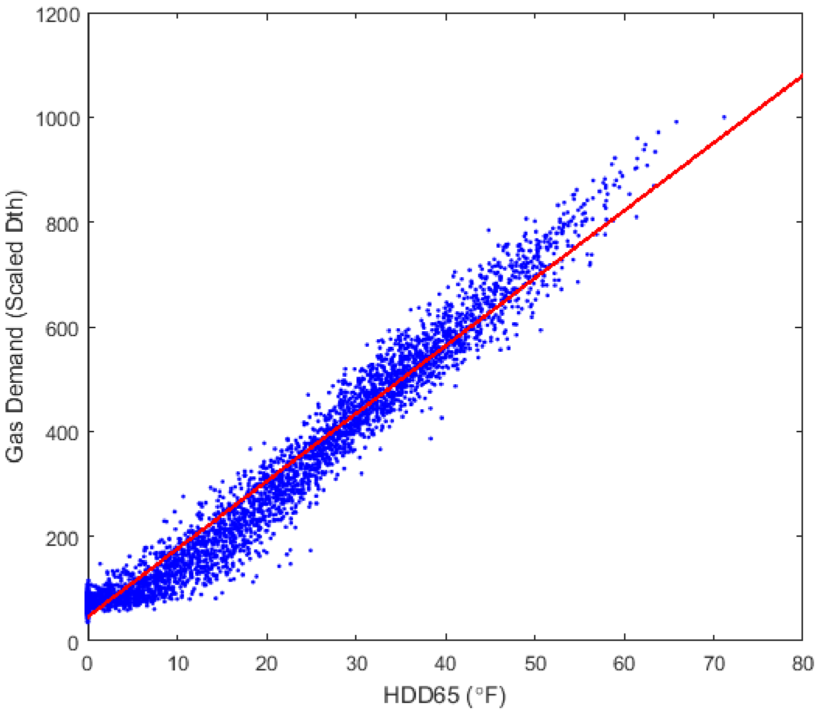


Figure 4.2: Historical daily gas demand against daily HDD. Reference temperature $T_{ref} = 65$. The red line shows demand linear with HDD

and population growth [47]. Exogenous and autoregressive terms such as prior day(s) temperature and demand also are used as predictors [35, 66].

Consider a linear regression model $\widehat{s} = \beta_0 + \sum_{i=1}^{m} \beta_i x_i$ that uses $m$ predictor variables ($x$) together with the historical demand data to estimate future values ($\widehat{s}$) of gas demand ($s$). The linear regression model assumes a linear relationship between the response variable (gas demand) and the predictor variables. Figures 4.1 and 4.2 suggest a linear relationship between $HDD$ and gas demand $s$ for temperatures below 65°F (positive $HDD$), which makes the linear assumption valid for the $HDD$ variable. If temperature change from the prior day

$$\Delta HDD = HDD_{\text{today}} - HDD_{\text{yesterday}}$$

is used as one of the predictors, the MLR model also makes an assumption about $\Delta HDD$'s linearity with gas demand. The exact relationship between $\Delta HDD$ and demand is unknown. Unlike $HDD$, the linear assumption is not necessarily valid for $\Delta HDD$. In fact, factors such as thermodynamic heat loss in buildings [30, 45] and behavioral responses [28, 30, 32] suggest a nonlinear relationship between $HDD$ change and gas demand.

If the nonlinear impact of prior day weather on demand were known, we could include this relationship into the linear regression model and improve our forecast accuracy. In this work, we investigate the relationship between temperature change and daily gas demand using Prior Day Weather Sensitivity ($PDWS$), an indicator variable introduced by Kaefer [31]. Analyzing the impact of prior day weather at different temperatures and different changes in temperature levels, we derive an equation $PDWS = f(HDD, \Delta HDD)$ relating $PDWS$ to $HDD$ and $\Delta HDD$.

A more robust linear regression model is developed that uses the result of the $PDWS$ impact factor. This new model is estimated from historical data, and we show that the $PDWS$-adjusted LR model yields improved forecasts.

The rest of this chapter is organized as follows. Section 4.2 discusses how gas distribution companies use the gas forecast in their operations planning process. In Section 4.3, we introduce common statistical models used in natural gas forecasting and highlight the input features that are relevant to this work. In Section 4.4, we investigate the impact of prior day weather on daily gas demand. The findings from Section 4.4 are used in Section 4.5 to develop a gas demand model that accounts appropriately for the prior day weather impact. We show in Section 4.5.2 the improvement in the forecast due to our contribution. In the concluding section, we discuss some of the implications of the results presented in this study. We also discuss how this result might also be used in design day studies. Finally, we offer recommendations about other variables (aside from temperature and change in temperature) that could be considered, including how this work can be extended to other temperature-driven demand such as electricity.

## 4.2 Operations Planning

In this section, we will discuss how gas demand forecasting is important to the efficient operation of gas utilities with an emphasis on the business-related risks associated with forecasting error. We also will consider “design day study” and “unusual days”, concepts that motivated this work.

Operational planning deals with the application of analytical methods in decision-making such that risk is minimized, while performance is maximized [68].

In the natural gas industry, plans regarding supply, distribution, and transmission of gas, including the design of gas infrastructure, are important to the efficient operation of the industry. Most of the planning efforts are based on the anticipated demand of end customers. Hence, natural gas utilities are highly motivated to forecast their customers' gas demand accurately.

### 4.2.1 Gas Nomination

In gas distribution companies, operations managers make short-term and long-term operation-critical decisions daily. Decisions regarding distribution, purchase, and storage of gas are made based on anticipated end customer demand, with the aim of minimizing both risk and operational cost. For instance, operations managers have to decide the exact quantity of natural gas to buy from the gas suppliers to meet customers' daily demand. This often is done by requesting a certain amount of gas supply in advance. The requested amount is called nomination. If the actual quantity of gas used turns out to be more or less than the nomination, the gas distribution utility is charged a penalty.

### 4.2.2 Flow Control Optimization

Gas control operators in utilities are charged with coordinating the flow of natural gas in the distribution network in a safe and efficient manner. The gas controllers, like the operations managers, need to know in advance the anticipated demand and the maximum expected flow to regulate the pressure and volume of gas in the transmission pipeline network. Inability to maintain the pressure and volume of gas in the distribution network can lead to a loss of service or to a pipeline explosion that may cause significant damage to lives and property [13].

### 4.2.3 Design Day

When building natural gas distribution infrastructure, utilities need to know how large their transmission pipelines need to be to handle peak demands. Since it is impossible to exceed the physical limit on the amount of gas that can be transmitted through a pipeline, the design of the natural gas transmission pipeline in a region is influenced by the highest probable demand for that region. This highest probable demand for which the system is designed is termed the design day demand, estimated through long-term forecasts [18].

### 4.2.4 Unusual Days

While accurate forecasting of daily demand is important to gas utilities, all days are not equally important. Some days have higher risk than others. For example, many utilities are not as concerned about forecasting during the summer compared to during the winter. During the winter, utilities are more worried about exceptionally cold days and days with unusual weather patterns than days of normal weather. We consider unusual day types including: coldest day in a year, colder than normal days, first cold days, first warm days, colder today than yesterday, and warmer today than yesterday. Most of the unusual days types are by definition sparse. Therefore, unusual days are not adequately represented in the historical data.

Models built to forecast gas demand use large historical data sets to predict future gas demand accurately. Since most of the critical days are not well represented in the historical data, they become difficult to forecast.

Given the physical and financial risks associated with these decisions, utilities are always interested in improving the accuracy of gas demand forecasts. In the next section, we will discuss techniques used in gas demand forecasting and factors that influence demand of gas.

### 4.3 Natural Gas Forecasting

Gas utilities employ forecasting experts with domain knowledge in natural gas forecasting. Gas forecasters develop statistical models to predict gas demand. The models typically are based on customers' demand history and factors such as weather (temperature, wind, dew point, and cloud cover) [47], seasonal factors (day of the week and holidays) [66], demographic (population and geographical location), and economic factors (price of gas and gross domestic product) [17]. Some of the common models include linear regression [26, 47, 66], nonlinear regression [5, 14, 67], artificial neural network [9, 34], fuzzy systems [9, 33], or an ensemble of two or more models [9, 33, 66].

### 4.3.1 Identifying Predictor-Response Relationships

When using a regression model, the relationship between the predictor and response variables must be known or approximated. Domain experts often provide good insight here. Models such as neural networks have the ability to infer predictor-response relationships [58]. However, neural networks are computationally intensive, and they are not good at extrapolating, unlike regression techniques.

In forecasting gas demand, linear regression is often used because of its simplicity and its ability to extrapolate. When using a regression model to forecast time series data, it is common to plot the response against each of the predictor variables. From this plot, the predictor-response relationship is inferred. We already showed in Figure 4.2 that the demand-*HDD* relationship is linear for natural gas. In this case, a linear regression model has the form

$$\widehat{s} = \beta_0 + \beta_1 HDD + \ldots .$$

Similar analysis is done for the other predictors to determine the nature of the relationship to gas demand. Sometimes, the relationship is simply assumed. For instance, a holiday effect can be taken as a linear factor or as an adjustment to day-of-the-week factors.

### 4.3.2 Other Temperature Variables

While many factors contribute to gas demand, most of the variance in the demand is explained by temperature variables such as $HDD$, $\Delta HDD$, and lagged temperature variables. We already identified the relationship between daily demand and $HDD$ as linear. In the next section, we will look more closely at $\Delta HDD$ and how it affects gas demand.

## 4.4 Impact of Prior Day Weather

Natural gas demand is highly correlated with temperature, but the relationship between temperature and demand is a complex one due to factors such as discomfort index [62], thermodynamics, and human behavioral response [16]. Daily gas demand depends not only on the day's weather condition, but also on the weather condition of the preceding day(s). For instance, a customer might use more gas than expected because the previous day was colder than normal. Hong [28] referred to a similar behavioral response in electric load forecasting as the recency effect. The term recency effect is extended from observations in psychology [25], where people tend to make decisions based on their most recent experience. Also, heat loss in buildings is a thermodynamic process, and the rate of heat loss depends on the nature of insulation system installed. Buildings with poor insulation tend to use more gas for space heating as they lose heat faster than those with good insulation.

### 4.4.1 Previous Investigations

Many studies have pointed out this nonlinear relationship [28, 35, 42, 66], but no study has considered the actual form of this nonlinearity. To account for this nonlinearity in natural gas models, load forecasting experts often employ lagged or moving average temperature variables [28, 35, 66]. The order of the lag variable often is obtained by trial and error or by analyzing the autocorrelation at various lags using the Akaike Information Criteria [42]. However, this technique is computationally intensive, since every lag has to be checked. While the lagging approach has been shown to produce better results than a single temperature component model, even better forecasting accuracy can be achieved if the nature of the nonlinearity is reflected in our models.

### Tenneti Index of Temperature Sensitivity

To understand the impact of temperature on gas demand forecast, Tenneti [61] introduced a temperature sensitivity index, which is a quantitative measure that ascribes a value from 0 to 1 to the contribution of temperature variables in a gas demand model. A value of 0 denotes the demand is not driven by temperature, while a value of 1 denotes the gas demand model is very temperature sensitive. In Tenneti's work, the temperature index is derived using only the daily temperature, and does not reflect the effect of prior day(s) weather conditions, but his work is easily extended to include the temperature-related variables of concern in our work.

### Prior Day Weather Sensitivity

Before the advent of forecasting software, some experienced gas forecasters predicted daily gas demand by drawing a line of best fit through historical data as illustrated in Figure 4.3. They were aware that the day's demand is influenced by the prior day temperature. Their intuition, based on experience and domain knowledge, was to calculate the difference between today's and yesterday's temperature and to go back one-third of the way from today's temperature. The adjusted temperature is used as the actual daily temperature, and the daily gas demand is estimated from the regression line. In Figure 4.3, if today's temperature is 40°F, the demand should be 570 units. To account for the impact of yesterday's

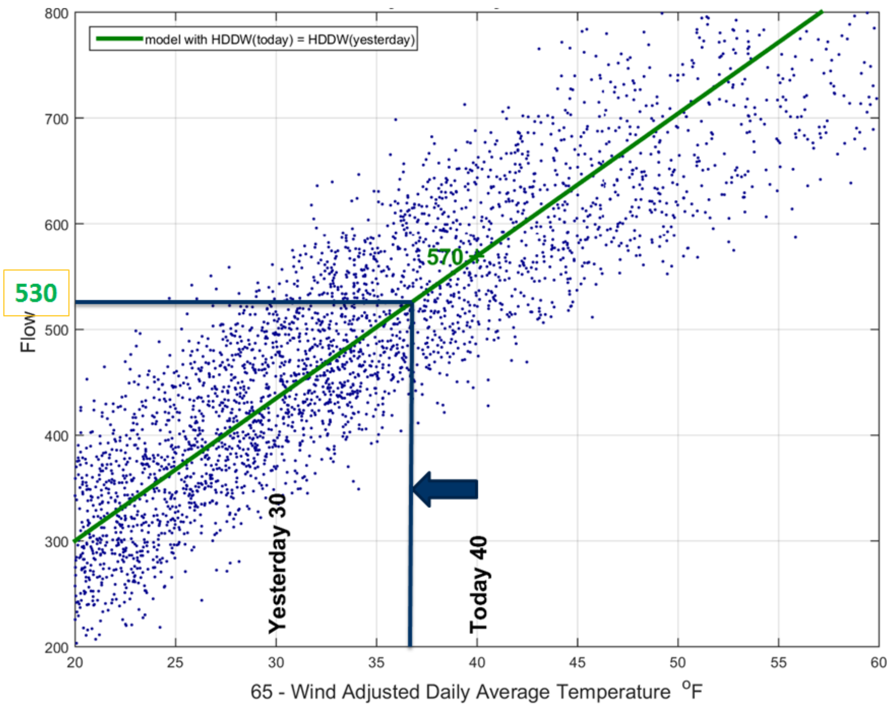


Figure 4.3: Adjusting gas estimate for prior day weather impact. Today is 40°F, and the demand estimate is 570 units, according to the regression line. But because yesterday was 10 °F warmer than today, gas demand is re-estimated as 530 units to account for prior day impact.

temperature, the experienced gas forecaster adjusted the demand to 530 units by moving down the regression line to a temperature of 37°F.

Kaefer [31] introduced the Prior Day Weather Sensitivity (PDWS). Similar to the Tenneti Index, the PDWS is quantitative metric that describes the impact prior weather conditions have on today's gas demand. The PDWS is estimated by fitting a three parameter linear regression model to historical temperature and consumption data. The three parameter model has current day temperature ($HDD_k$) and change-in-temperature from previous day ($\Delta HDD_k$) as the independent features and actual gas consumption ($s_k$) as the dependent variable. The model coefficients are evaluated, and the *PDWS* is calculated,

$$\widehat{s}_k = \beta_0 + \beta_1 HDD_k + \beta_2 \Delta HDD_k \text{ , and} \tag{4.1}$$

$$PDWS = -\frac{\beta_2}{\beta_1} \ . \tag{4.2}$$

Evaluating this impact factor using historical data collected from more than 150 operating areas across the USA, Kaefer [31] showed that for most temperature-dependent operating areas, the PDWS is between $-0.3$ and $-0.2$. Those values are highly correlated with expert knowledge mentioned previously. In our framework, the gas forecasters had chosen a PDWS of $-\frac{1}{3} \approx 0.3$.

### 4.4.2 Further Investigations

The PDWS proposed by Kaefer only tells us the impact factor for an operating area and offers no insight as to whether the effect differs at different temperature levels. We do not know if the prior day impact is higher or lower during bitter cold temperatures. Our intuition suggests that people might act differently during a bitter cold weather or when there is a large swing in temperature. Using a constant value of PDWS for all temperature ranges would be inaccurate if this hypothesis is correct.

We investigate the Prior Day Weather Sensitivity further to explore the impact of prior day weather on daily gas demand and to provide answers to questions such as:

- Does PDWS vary by any independent variable?
- How does this independent variable affect the PDWS?

The methods used in this paper to determine the PDWS are an extension of Kaefer [31]. For this study, we examine the Prior Day Weather Sensitivity at different values of temperature and change in temperature.

### By Temperature

**Data:** For the analysis in this and the following sections, we use daily natural gas demand and the corresponding temperature time series data from 2003 to 2015 for several regions in the United State. The data obtained from actual gas utilities have been anonymized to protect the identity of the gas utilities. The data is divided into training and testing sets. Unless otherwise stated, the training set includes 2003 to 2012 data, and the test set includes 2013 to 2015 data. The temperature variables have been transformed to HDD.

**Procedure:** To determine the prior day impact by temperature, we partition the training data by temperature ranges. We are interested in determining the prior day impact for only the heating days, so we discard all data points in the training data with $HDD = 0$ (non-heating days). For each partition, the PDWS is calculated from the parameters of Equation (4.2) obtained by fitting the linear regression in Equation (4.1) to the data points in that partition. This is done by sorting the time series in ascending order of temperature and sliding through the sorted data using a rectangular window of length $l$ and lag $\tau$. For each window, the temperature, prior day temperature, and the corresponding gas consumption are used to build the model in Equation (4.1), and the PDWS is calculated. Figure 4.4 shows the PDWS obtained against the average *HDD* in each window. For Figure 4.4, we used a window length of 500 and a lag of 50.

**Result:** The result presented in Figure 4.4 provides us with new insight about the nature of the impact of prior day weather on daily gas demand. Figure 4.4 shows that PDWS changes by temperature as opposed to being constant, as suggested by expert knowledge and prior study by Kaefer et al. [31]. The figure also tells us how PDWS changes by temperature. Figure 4.4 suggests a decaying exponential relationship between PDWS and HDD, with the impact of prior day weather higher at lower HDD (warmer temperatures) and lesser at higher HDD (colder temperatures). Using an exponential function of the form

$$PDWS = \gamma_0 + \gamma_1 \cdot e^{\gamma_2 \cdot HDD} \ , \tag{4.3}$$

Equation 4.3 was fitted to the PDWS vs. HDD curve (blue line). Using nonlinear least square regression, we estimate the optimum parameters of the exponential function. The optimal parameters of the exponential function are used to estimate the PDWS given HDD as shown by the green line in Figure 4.4. The PDWS for was estimated for larger values of the HDD (not available the training data) to see how well the exponential function extrapolates. As seen in Figure 4.4, the PDWS levels off at around $-0.1$ as HDD approaches large values.

### By Temperature and Change in Temperature

We extend the investigation in Section 4.4.2 by assessing the impact of prior day temperature at different temperature changes. Having obtained evidence that the PDWS varies by temperature, we want to know if the impact is different with large or small temperature swings.

**Data** The same data from the previous analysis is used.

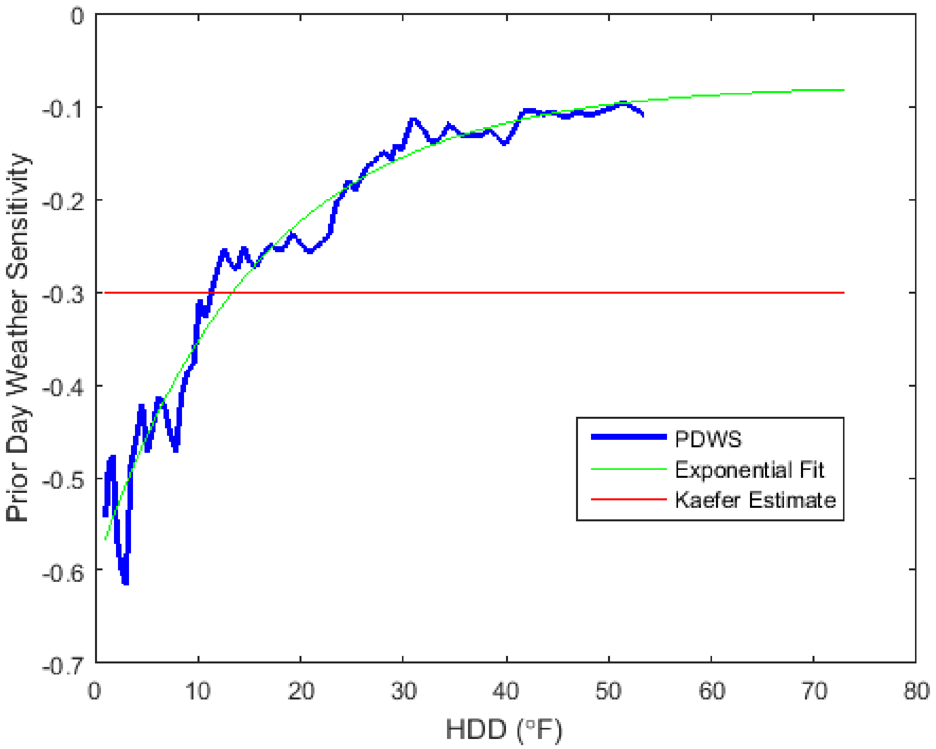


Figure 4.4: Prior Day Weather Sensitivity varying by temperature. The *PDWS* plot suggests *PDWS* varies exponentially with temperature.

**Procedure** First, we partition the training data by temperature as in Section 4.4.2. Each temperature partition is further partitioned by change-in-temperature from the previous day. The PDWS is calculated the same way as in Section 4.4.2 from the data in each inner partition. The PDWS is plotted against the average of the temperatures and the average of the change in temperature. The surface plot obtained is shown in Figure 4.5.

**Result:** The surface plot shows that the change in temperature also influences the impact of prior day temperature. From the surface plot, we observe that the

$PDWS$ is close to zero at one extreme end of the surface (high $HDD$ and positive $\Delta HDD$), which suggests the impact of prior day weather is less for this weather condition. The impact is more at the low values of $HDD$ and negative $\Delta HDD$. Although there is high variance in the underlying data, looking at the $PDWS$ surface from the $HDD$ axis, the exponential form can still be observed. On the $\Delta HDD$ axis, there seems to be a linear relationship between $PDWS$ and $\Delta HDD$. As $\Delta HDD$ goes from negative to positive, the $PDWS$ increases i.e. the impact of prior day weather is lesser when there is a positive change in temperature from prior day than when the temperature change is negative.

To estimate the $PDWS$ as a function of $HDD$ and $\Delta HDD$, we fit a nonlinear model of the form

$$PDWS = \gamma_0 + \gamma_1 \cdot e^{\gamma_2 \cdot HDD} + \gamma_3 \cdot \frac{1 - e^{\gamma_4 \cdot \Delta HDD}}{1 + e^{\gamma_4 \cdot \Delta HDD}} \ , \tag{4.4}$$

to the $PDWS$ surface. The $HDD$ component of Equation 4.4 models the exponential relationship between $PDWS$ and $HDD$. For the $\Delta HDD$ component, we use a logistic sigmoid function rather than a linear function. The $PDWS$ surface in Figure 4.5 suggest a linear relationship between $PDWS$ and $\Delta HDD$. However, we know from domain knowledge that the $PDWS$ can not be greater than zero. Using a linear function, for large positive values of $\Delta HDD$, the model will extrapolate and estimate $PDWS$ greater than zero. The same is true for large negative values. A linear $\Delta HDD$ component will extrapolate to unreasonably small values of $PDWS$. For this reason, we use a logistic sigmoid function described in Equation 4.4. This sigmoid function ensures that the $PDWS$ - $\Delta HDD$ relationship is near linear for values of $\Delta HDD$ within a bound, and levels out as $\Delta HDD$ goes outside that boundary. Equation 4.4 ensures that the $PDWS$ model extrapolates well for extreme values of $HDD$ and $\Delta HDD$.

The coefficient $\gamma$ of the $PDWS$ model is evaluated through nonlinear optimization. The value of $\gamma$ that minimizes the sum squared error between the estimate of $PDWS$ (Equation 4.4) and actual value of $PDWS$ is returned as the optimum $\gamma$. Figure 4.6 shows the fitted surface. The surface fit is the estimate of $PDWS$ for different combinations of $HDD$ and $\Delta HDD$. It can be seen how the model extrapolates for large positive and negative values of $\Delta HDD$.

### 4.5 Accounting for PDWS in Demand Forecast

The analysis in Section 4.4 provides us with new information about the nature of the prior day weather sensitivity and how it impacts gas demand. In this section, we incorporate this finding into a gas forecasting model to predict demand. In the last section, we estimated $PDWS$, first as a function of only $HDD$, then as a function of both $HDD$ and $\Delta HDD$. In this section, we build gas demand regression

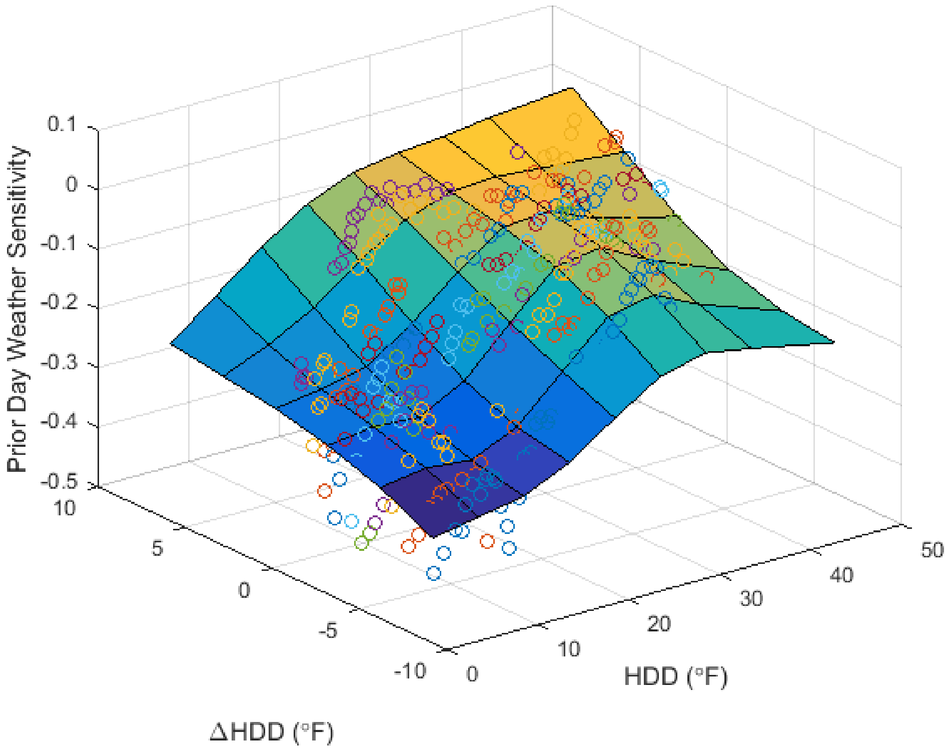


Figure 4.5: Prior Day Weather Sensitivity vs. temperature and change-in-temperature. The *PDWS* surface suggest that *PDWS* varies linearly with temperature-change in addition to varying exponentially with temperature.

models based the two *PDWS* functions derived in Section 4.4.2. The regression models use the estimate of *PDWS* as one of the independent variables.

### 4.5.1 Gas Demand Forecast - Model Description

Four simple linear regression models were built to evaluate the contribution of the *PDWS* discussed in Section 4.4.2 on gas demand. All four models are temperature-only models. The models do not account for trends [19, 29] in the historical data, nor are seasonal factors such as day of the week accounted for. The

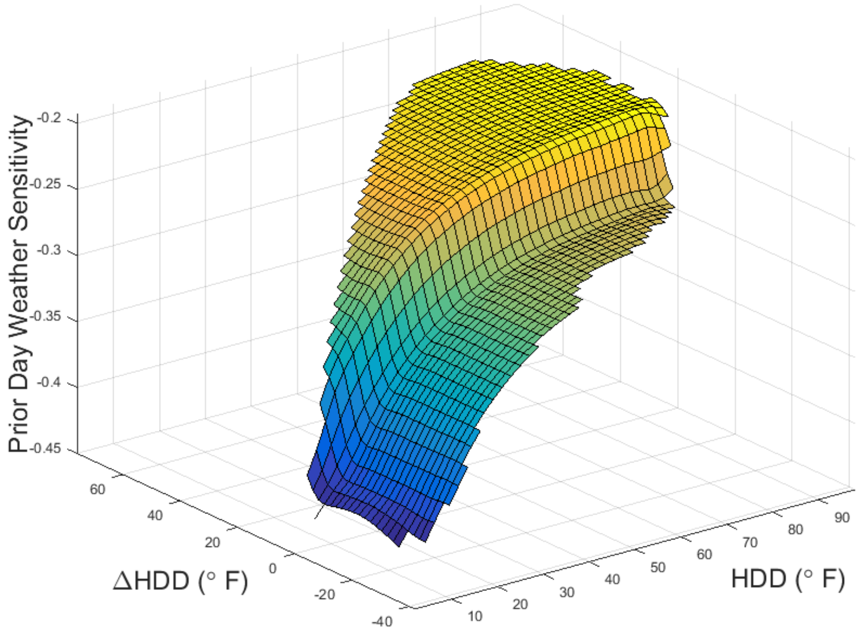


Figure 4.6: Prior Day Weather Sensitivity surface fit. *PDWS* is an exponential function along the *HDD* axis and a sigmoid function along the $\Delta HDD$ axis.

temperature-only models were used so the impact of the temperature can be studied in isolation. Model #1 (Kaefer) is a three parameter LR model. This model is based on the assumption that the *PDWS* is a constant factor according to Kaefer [31]. Model #2 (Linear) is a four-parameter model based on the discussion in Section 4.4.2. For this model, it is assumed the *PDWS* varies linearly with temperature. Model #3 (Exponential) is a five-parameter model based on *PDWS* being a function of temperature only. For this model, the *PDWS* derived in Equation (4.3) is used. Model #4 (Exponential-Sigmoid) is also a five parameter model like Model #3 except that its *PDWS* is a function of both temperature and change in temperature. The *PDWS* function derived in Equation (4.4) is used in this model. Let $\widehat{s}_k$ represent the estimate of natural gas demand for day $k$; $HDD_k$ represent the corresponding Heating Degree Day, and $\Delta HDD_k$ represent the change in Heating Degree Day from the previous day.

$$\#1 - \text{Kaefer}: \ \widehat{s}_k = \beta_0 + \beta_1 HDD_k + \beta_2 \Delta HDD_k \ , \tag{4.5}$$

$$\#2 - \text{Linear}: \ \widehat{s}_k = \beta_0 + \beta_1 HDD_k + \beta_2 \Delta HDD_k + \beta_3 HDD_k \cdot \Delta HDD_k \ , \tag{4.6}$$

$$\begin{aligned} \#3 - \text{Exponential}: \ \widehat{s}_k = \beta_0 + \beta_1 HDD_k + \beta_2 \Delta HDD_k + \beta_3 HDD_k \cdot \Delta HDD_k \\ + \beta_4 PDWS\left(HDD_k\right) \cdot \Delta HDD_k \ , \end{aligned} \tag{4.7}$$

$$\begin{aligned} \#4 - \text{Exponential-Sigmoid}: \ \widehat{s}_k = \beta_0 + \beta_1 HDD_k + \beta_2 \Delta HDD_k + \beta_3 HDD_k \cdot \Delta HDD_k \\ + \beta_4 PDWS\left(HDD_k, \Delta HDD_k\right) \cdot \Delta HDD_k \ . \end{aligned} \tag{4.8}$$

### 4.5.2 Performance Evaluation

The models in Equations (4.5) to (4.8) are trained on ten years of historical demand and temperature data from 2003 to 2012. The remaining three years of data (2013 to 2015) were held out for testing. The performance of Models #1, #2, #3 and #4 were evaluated on the test data. For this evaluation, Root Mean Square Error (RMSE) is used as error metric. In Section 4.2.4, we highlighted unusual days as one of the motivations for this work. For that reason, we compare all models based on their performance on the various unusual days types.

First, the Exponential model (#3) is compared against Kaefer (#1) and the Linear model (#2). The Exponential-Sigmoid model (#4) is also compared against Models #1 and #2. Figure 4.7 shows the RMSE obtained on the testing set, plotted by unusual day types (see Section 4.2.4). The bars represent the categories of unusual days. The bar chart contains four groupings, each representing the forecast error for Kaefer, Linear, Exponential, and Exponential-Sigmoid models. In Table 4.1, we highlight a few of the unusual days and compares the RMSE of the Exponential (#3) against that of the Exponential-Sigmoid (#4) models. The '% $\Downarrow$' columns represent the percentage reduction in RMSE between the Exponential, Exponential-Sigmoid models and the Kaefer model i.e.
$\#3\ (\%\Downarrow) \implies \frac{Model\#1\ RMSE\ -\ Model\#3\ RMSE}{Model\#1\ RMSE} * 100\%$ and
$\#4\ (\%\Downarrow) \implies \frac{Model\#1\ RMSE\ -\ Model\#4\ RMSE}{Model\#1\ RMSE} * 100\%$. Comparing the performance on the unusual day types shown in Table 4.1, the Kaefer model (#1) has the highest RMSE for most day types, while #3 and #4 has the least RMSE. For the first five unusual day types highlighted, both the Exponential (#3) and Exponential-Sigmoid (#4) models show reduction in forecast error. For the 'Warmer today than yesterday' however, we got increment in forecast error for both models #3 and #4.

So far, we have based our analysis on dataset from one gas utility. To assert the consistency of our result, we perform the same test on dataset from other gas utilities. Data from three other gas utilities was obtained in addition to the one used in earlier evaluations. The same range of data as in the previous evaluation was used. Data from 2003 to 2013 were used as training set, 2013 to 2015 data used for testing. Tables 4.2 and 4.3 shows the percentage reduction in RMSE between the Kaefer, the Exponential, and Exponential-Sigmoid models for four datasets obtained from different gas utilities. In Table 4.2, we report the percentage reduction in RMSE for four operation areas (Op) using the Exponential model. While, Table 4.3 shows the percentage reduction in RMSE due to the Exponential-Sigmoid model. Looking at the results presented in Tables 4.2 and 4.3 we can see a consistent reduction in forecast error for all unusual day types except for the 'Warmer today than yesterday' day type.

The most significant performance improvement was observed for 'Colder today than yesterday' days with up to 25% reduction in RMSE for the Exponential model. The only instance in which our Exponential models performed worse than the Kaefer model is 'Warmer today than yesterday'. The same is true for the Exponential-sigmoid model, with up to 23% reduction in RMSE on 'Colder today than yesterday' and −22% on 'Warmer today than yesterday'. Adjusting for the prior day weather sensitivity on 'Warmer today than yesterday' day type did not improve the forecast accuracy but instead increased the forecast error. For most of the other unusual days, our contribution (which is the Kaefer model adjusted for prior day impact) lead to reduction in forecast error.

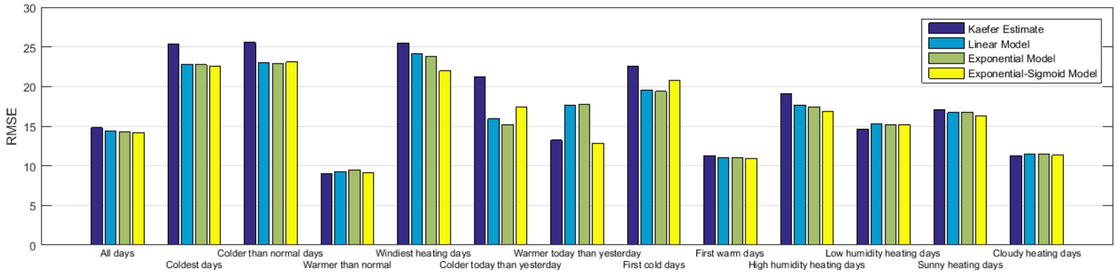


Figure 4.7: Gas demand forecast RMSE by unusual days. The four bars represent each of the four models compared side-by-side. The Exponential and Exponential-Sigmoid models performed better than the Kaefer and Linear models for most of the unusual days type identified.

Table 4.1: Models Performance by Unusual days. The RMSE of the four models are being compared. Model 3 (Exponential) and Model 4 (Exponential-Sigmoid) both have lower RMSE than Model 1 (Kaefer Estimate) and Model 2 (Linear) for most of the unusual days type.

| | RMSE (Dth) | | | | | |
|---|---|---|---|---|---|---|
| Unusual day types | # 1 | #2 | #3 | #3 (% ⇓) | #4 | #4 (% ⇓) |
| Coldest day | 5.76 | 5.31 | 5.25 | 9.0 | 5.31 | 7.9 |
| Colder than normal | 5.32 | 4.74 | 4.66 | 12.3 | 4.74 | 10.7 |
| Warmer than normal | 1.72 | 1.50 | 1.54 | 10.5 | 1.50 | 13.1 |
| Colder today than yesterday | 5.00 | 3.86 | 3.73 | 25.5 | 3.85 | 23.1 |
| First cold days | 2.90 | 2.53 | 2.50 | 13.8 | 2.53 | 12.5 |
| Warmer today than yesterday | 3.02 | 3.72 | 3.80 | -26.1 | 3.63 | -22.0 |

## 4.6 Future Work

So far, we have considered the impact of prior day weather on daily gas demand.

We investigated Prior Day Weather Sensitivity and explored two variables on which $PDWS$ depends. We examined $PDWS$ by temperature ($HDD$) and temperature change ($\Delta HDD$), and we derived $PDWS$ as a function of $HDD$ (Equation 4.3) and as a function of both $HDD$ and $\Delta HDD$ (Equation 4.4). By adjusting a

Table 4.2: Percentage reduction in RMSE resulting from the Exponential model. This model took into account *PDWS* varying exponentially with temperature. The performance is compared across four gas utilities.

| | Exponential (#3) - RMSE ⇓ (%) | | | | |
|---|---|---|---|---|---|
| Unusual day types | Op 1 | Op 2 | Op 3 | Op 4 | Average |
| All days | 1.94 | 0.98 | 1.27 | 1.33 | 1.38 |
| Coldest days | 8.97 | 6.87 | 6.21 | 4.19 | 6.56 |
| Colder than normal | 12.31 | 5.5 | 7.47 | 5.36 | 7.66 |
| Warmer than normal | 10.5 | -0.56 | 0.45 | -3.94 | 1.61 |
| Windiest heating days | 8.28 | -2.11 | 19.35 | 7.66 | 8.30 |
| Colder today than yesterday | 25.51 | 19.99 | 20.15 | 19.19 | 21.21 |
| Warmer today than yesterday | -26.05 | 26.04 | -14.18 | -26.81 | -23.27 |
| First cold days | 13.77 | 9.55 | 10.33 | 8.98 | 10.66 |
| First warm days | 2.89 | 1.09 | 1.46 | 1.5 | 1.74 |
| High humidity heating days | 4.59 | 10.15 | 14.24 | 8.07 | 9.26 |
| Low humidity heating days | 4.72 | -9.04 | 0.83 | -3.79 | -1.82 |
| Sunny heating days | -1.45 | -2.18 | 1.87 | 3.54 | 0.45 |
| Cloudy heating days | 1.1 | -1.54 | 0.63 | 0.93 | 0.28 |

three-parameter linear regression model (Kaefer model) with the *PDWS* (from Equations 4.3 and 4.4) to arrive two five-parameter LR models (Exponential and Exponential-Sigmoid model), we showed in Section 4.5.2 that both Exponential-*PDWS* models offers significant improvement over the Kaefer model, especially for the cold unusual days types.

Looking at Tables 4.2and 4.3, we see that our prior-day adjusted models consistently perform worse than a basic LR model on 'Warmer today than yesterday' day type. Considering the *PDWS* surface in Figure 4.5, 'Warmer today than yesterday' implies that the $\Delta HDD$ is less than zero, so that points in this space make up the 'Warmer today than yesterday' day type. A 'Colder today than

Table 4.3: Percentage reduction in RMSE resulting from the Exponential-Sigmoid model. The Exponential-Sigmoid model accounts for *PDWS* being a function of both temperature and temperature-change. The performance is compared across four gas utilities.

| | Exponential-Sigmoid (#4) - RMSE ⇓ (%) | | | | |
|---|---|---|---|---|---|
| Unusual day types | Op 1 | Op 2 | Op 3 | Op 4 | Average |
| All days | 1.88 | 1.07 | 1.27 | 1.33 | 1.39 |
| Coldest days | 7.91 | 8.46 | 5.45 | 5.17 | 6.75 |
| Colder than normal | 10.74 | 5.40 | 5.46 | 5.06 | 6.67 |
| Warmer than normal | 13.11 | 2.30 | 1.52 | -1.51 | 3.86 |
| Windiest heating days | 7.86 | 0.34 | 16.81 | 8.06 | 8.27 |
| Colder today than yesterday | 23.1 | 18.39 | 16.78 | 12.07 | 17.59 |
| Warmer today than yesterday | -22.01 | -19.25 | -4.04 | -12.27 | -14.39 |
| First cold days | 12.53 | 8.07 | 8.75 | 6.19 | 8.889 |
| First warm days | 2.80 | 0.00 | 1.68 | 1.59 | 1.52 |
| High humidity heating days | 4.57 | 12.55 | 12.47 | 5.40 | 8.75 |
| Low humidity heating days | 6.43 | -8.01 | 2.11 | -3.79 | -0.82 |
| Sunny heating days | -1.49 | -1.78 | 1.12 | 3.86 | 0.43 |
| Cloudy heating days | 1.64 | -0.66 | 0.69 | -0.56 | 0.28 |

yesterday' implies that the $\Delta HDD$ is greater than zero. Since we obtained evidence that the our adjustment works for 'Colder today than yesterday' but not 'Warmer today than yesterday', introducing a 'kink' into the gas demand model at $\Delta HDD = 0$ such that the model has components for both $\Delta HDD < 0$ and $\Delta HDD > 0$, might offer a performance improvement for 'Warmer today than yesterday' day type.

In Section 4.4.2, based on the calculated values of *PDWS*, we used an exponential function (Equation 4.3) to relate *HDD* to *PDWS*. We know that our exponential function estimates better than a linear function. Other rational functions that fit the data can also be used. However, care must be taken in ensuring the rational function has reasonable asymptotes. For any chosen rational function, the horizontal asymptote should describe the behavior of $PDWS$ as $HDD$ gets large ensuring the $PDWS$ value stays within a reasonable bound.

### 4.6.1 Investigating other variables

Our analysis, has shown that $PDWS$ varies by $HDD$ and $\Delta HDD$. As recommendation for future work, other variables can be explored using the same method to determine if and how they affect the $PDWS$. For instance, if the prior day is an holiday, would that impact the prior day weather sensitivity? If it does, would adjusting for that factor offer any significant improvement to the forecast accuracy.

In this chapter, we only looked at one prior day weather impact on gas demand. We could also considered the impact of prior $x$ days weather conditions on daily gas demand. Earlier in Section 4.4.1, we mentioned that higher-order lagged temperature variables often are used in gas forecasting models. If using a linear regression model, the coefficient of the lagged temperature is assumed to be constant with demand. Prior $x$ days could be investigated the same way prior one day has been investigated in this chapter.

### 4.6.2 Application in other temperature-driven demand

Like natural gas demand, electric load depends on largely on temperature [21, 22, 24, 28, 48]. Hence, models used in forecasting electric load are similar to gas demand models. In fact, electricity and natural gas are used interchangeably for space heating. Most models used to forecast electric load also use lagged temperature variables [20, 27, 28, 35] to account for the complex interaction between temperature and electric load. The work presented in this chapter can be extended to electric load forecasting.

We hypothesized in Section 4.4 that the complex relationship between temperature and demand is due to a possible behavioral response [16, 62] or recency effect [28], where people make decisions based on recent experience [25], such as a customer using more gas simply because the previous day was very cold. In this regard, our work can be applicable in financial forecasting, predicting customer's buying patterns, or any other predictive analysis where recency effects can be observed.

## 4.7 Conclusion

In this chapter, we presented a method by which the accuracy of natural gas demand forecast can be improved. We highlighted a number of reasons why an accurate forecast is important to gas utilities. We also pointed out certain days (such as design day and unusual days) when an accurate forecast is especially important. By investigating the impact of prior day weather conditions on daily gas demand using a metric referred to as ‘Prior Day Weather Sensitivity’, we derived $PDWS = f(HDD)$ and $PDWS = f(HDD, \Delta HDD)$ (see Equations 4.3 and 4.4), showing that the impact of prior day weather depends on both the current day temperature and temperature change from the previous day. Two five-parameter linear regression models (adjusted for the prior day impact factor) were estimated from historical gas demand and temperature data. The performance of the (prior-day adjusted) models were compared to two other linear regression models - one assumes a constant $PDWS$, while the other assumes $PDWS$ varies linearly with temperature. We showed in Tables 4.1, 4.2, and 4.3 that our prior day adjustment improves the gas estimate with up to 25% reduction in RMSE.

# CHAPTER 5

## Research Contributions and Recommendations

This chapter provides a summary of the challenges addressed, ideas expressed, and techniques employed throughout this work. We provide a recap of the results from Chapters 2, 3, and 4, stating their contributions to gas demand forecasting and their value to gas utilities. We offer recommendations on how the ideas presented in this thesis can be used, extended, or applied to related fields of forecasting.

### 5.1 Research Contributions

In this thesis, we explored different approaches by which the accuracy of gas demand forecast can be improved, especially on days that pose major risks to gas utilities. The techniques discussed in this thesis are focused on achieving better forecasts on days with unusual weather patterns, during periods of extreme cold events. We explored the impact of temperature and human behavioral effects on daily gas demand and provided means by which their complex interactions can be incorporated into a gas demand model.

In Chapter 2, we presented a semi-supervised pattern recognition algorithm to identify extreme cold events in natural gas demand data. We pointed out an unusual response in gas demand to temperature during periods of extreme cold weather and considered why the unusual dynamics is a challenge in forecasting gas demand. We demonstrated that by performing a low-dimensional embedding of the gas demand data, we could cluster natural gas data based on similarities in dynamics. We showed that our RPS-$k$NN algorithm was able to identify extreme cold events in the historical gas data. The extreme cold events identified in this chapter were used in Chapter 3 in an effort to improve gas forecasts during extreme cold events.

In Chapter 3, we discussed a strategy by which the accuracy of gas demand forecasts can be improved during periods of extreme cold events. We showed a characteristic pattern in the forecast residuals for days in an extreme cold event, and postulated an unmodeled behavioral component. We presented a residual learning architecture to learn the statistics of the residuals for days in the extreme events identified by our RPS-$k$NN algorithm in Chapter 2. The estimates of the residuals produced by the residual model were used to adjust an initial forecast. The performance of the adjustment model was evaluated using the Mean Absolute Percentage Error between the new estimate and actual demand. The adjustment model's MAPE was compared to the initial base model estimate. Our adjustment model performed better for some days in the identified events and performed worse on other days. There was no conclusive evidence that our residual learning technique would improve forecast accuracy during future extreme cold events.

In Chapter 4, we presented a technique to improve the accuracy of gas forecasts during days with extreme and unusual weather patterns. We outlined certain day types as unusual days, either because those days are challenging to forecast (gas demand) or have high business-related risk. We highlighted the complex relationship between prior day temperature and daily demand, and analyzed the impact of prior day temperate using *PDWS*. A previous study [31] evaluated the *PDWS* as constant. We showed that the *PDWS* depends on both temperature and change-in-temperature. Using the *PDWS* function we derived, we developed a gas demand forecast model that uses our *PDWS* factor. The performance of our *PDWS*-adjusted model was compared to other models (without the *PDWS* adjustment). We showed a performance improvement due to our *PDWS* adjustment. For instance, for the 'Colder today than yesterday' day type, we achieved up to 25% reduction in RMSE.

## 5.2 Recommendations

In Chapter 2, we identified extreme cold events for the purpose of learning the statistics of their forecast residuals. While we focused our attention of identification of certain extreme cold events, the technique developed in this thesis can be extended to classifying extreme cold events. We considered only five-day events in our analysis. However, some extreme cold events could be three or seven-day event. The RPS-$k$NN algorithm can be re-purposed for classification task as it can identify events in similar class based on their flow-temperature dynamics.

The prior day weather sensitivity equation derived in Chapter 4 can be used in design day studies. In estimating design day gas demand, the design day temperature and prior day temperature are used. D'Silva in [23] developed an algorithm to determined the design day temperature by estimating the 1-in-$n$ years temperature using a nonparametric distribution. The prior design day temperature is derived from the 1-in-$n$ temperature and the $PDWS$ factor (previously determined as constant by [31]). In this study, we have shown that the $PDWS$ is not constant, but varies by temperature. Using the $PDWS$ function derived in Chapter 4, we can obtain a better estimate of prior design day temperature and improve the design day gas estimate.

In [31], Kaefer developed a surrogate data selection algorithm to supplement areas with insufficient data. Kaefer's surrogate algorithm uses $PDWS$ [31] to select a subset of donors from a pool of donors. Kaefer's surrogate selection is based on constant $PDWS$. We have shown in Chapter 4 that $PDWS$ varies with temperature and temperature-change. The surrogate selection algorithm should be modified to account for the varying $PDWS$.

## 5.3 Concluding Remarks

This thesis explored methods by which the accuracy of gas demand forecast can be improved during the hard-to-forecast days. Throughout this thesis, we have shown the unusual dynamics of extreme cold events. In Chapter 2, we demonstrated how extreme events can be identified using pattern recognition algorithms. In Chapter 3, we demonstrated how, by learning of forecast residuals during extreme cold events, the accuracy of forecasts can be improved. In Chapter 4, we showed that the impact of prior day weather on daily gas demand is not a constant factor but varies with temperature and temperature-change. We demonstrated how by appropriately accounting for the prior day impact factor, we improved the accuracy of gas forecasts for most of the unusual day types.

# References

[1] H. Abarbanel, *Analysis of Observed Chaotic Data.* Springer Science and Business Media, 2012.

[2] H. Abdi, "Partial least square regression (PLS regression)," *Encyclopedia for Research Methods for the Social Sciences*, pp. 792–795, 2003.

[3] American Petroleum Institute, "Natural Gas and Its Uses," 2015, accessed: 2016-02-16. [Online]. Available: http://www.api.org/Oil-and-Natural-Gas-Overview/Exploration-and-Production/Natural-Gas/Natural-Gas-Uses

[4] H. Aras and N. Aras, "Forecasting residential natural gas demand," *Energy Sources*, vol. 26, no. 5, pp. 463–472, 2004.

[5] N. Aras, "Forecasting residential consumption of natural gas using genetic algorithms," *Energy, Exploration and Exploitation*, vol. 26, no. 4, pp. 241–266, 2008.

[6] M. Ardalani-Farsa and S. Zolfaghari, "Chaotic time series prediction with residual analysis method using hybrid Elman–NARX neural networks," *Neurocomputing*, vol. 73, no. 13, pp. 2540–2553, 2010.

[7] M. Ardalani-Farsa and S. Zolfaghari, "Residual analysis and combination of embedding theorem and artificial intelligence in chaotic time series forecasting," *Applied Artificial Intelligence*, vol. 25, no. 1, pp. 45–73, 2011. [Online]. Available: http://dx.doi.org/10.1080/08839514.2011.529263

[8] G. Arduino, P. Reggiani, and E. Todini, "Recent advances in flood forecasting and flood risk assessment," *Hydrology and Earth System Sciences Discussions*, vol. 9, no. 4, pp. 280–284, 2005.

[9] A. Azadeh, S. Asadzadeh, and A. Ghanbari, "An adaptive network-based fuzzy inference system for short-term natural gas demand estimation: Uncertain and complex environments," *Energy Policy*, vol. 38, no. 3, pp. 1529–1536, 2010.

[10] A. Basharat and M. Shah, “Time series prediction by chaotic modeling of nonlinear dynamical systems,” in *Computer Vision, 2009 IEEE 12th International Conference on.* IEEE, 2009, pp. 1941–1948.

[11] D. A. Belsley, E. Kuh, and R. E. Welsch, *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.* John Wiley & Sons, 2005, vol. 571.

[12] J. Black, “Cost and performance baseline for fossil energy plants volume 1: Bituminous coal and natural gas to electricity,” *Final report (2nd ed.) National Energy Technology Laboratory (2010 Nov) Report no.: DOE20101397*, 2010.

[13] Black Hills Energy, “PIPELINE SAFETY. What can you do?” 2011, accessed: 2016-02-28. [Online]. Available: http://www.kcc.state.ks.us/pipeline/2011_seminar/black_hills_safety_presentation.pdf

[14] M. Brabec, O. Konár, E. Pelikán, and M. Malý̀, “A nonlinear mixed effects model for the prediction of natural gas consumption by individual customers,” *International Journal of Forecasting*, vol. 24, no. 4, pp. 659–678, 2008.

[15] R. H. Brown, “Research results: The heck-with-it hook and other observations,” in *Southern Gas Association Conference: Gas Forecasters Forum*, October 16 2007.

[16] ——, “In search of the hook equation: Modeling behavioral response during bitter cold events,” in *2014 Gas Forecasters Forum*, October 16 2014.

[17] R. H. Brown, D. Clark, G. F. Corliss, F. Nourzad, T. Quinn, and C. Twetten, “Forecasting natural gas demand: The role of physical and economic factors,” in *32nd Annual International Symposium on Forecasting*, 2012. [Online]. Available: http://www.forecasters.org/proceedings12/QUINN_THOMAS_ISF2012_2012-07-16.pdf

[18] R. H. Brown, P. E. Kaefer, C. R. Jay, and S. R. Vitullo, “Forecasting natural gas design day demand from historical monthly data,” in *Proceedings of the Pipeline Simulation Interest Group 2014 Conference*, May 6-9 2014.

[19] R. H. Brown, S. R. Vitullo, G. F. Corliss, M. Adya, P. E. Kaefer, and R. J. Povinelli, “Detrending daily natural gas consumption series to improve

short-term forecasts," in *IEEE Power and Energy Society 2015 Conference*, July 26-30 2015.

[20] C. Crowley and F. L. Joutz, "Weather effects on electricity loads: Modeling and forecasting 12 December 2005," *Final Report for US EPA on Weather Effects on Electricity Loads*, 2005, accessed: 2016-03-22. [Online]. Available: http://www.ce.jhu.edu/epastar2000/epawebsrc/joutz/Final%20Report%20EPA%20Weather%20Effects%20on%20Electricity%20Loads.pdf

[21] H. Cui and X. Peng, "Short-term city electric load forecasting with considering temperature effects: An improved ARIMAX model," *Mathematical Problems in Engineering*, vol. 2015, 2015.

[22] A. P. Douglas, A. M. Breipohl, F. N. Lee, and R. Adapa, "The impacts of temperature forecast uncertainty on Bayesian load forecasting," *Power Systems, IEEE Transactions on*, vol. 13, no. 4, pp. 1507–1513, 1998.

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

[24] E. E. Elattar, J. Y. Goulermas, and Q. H. Wu, "Electric load forecasting based on locally weighted support vector regression," *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, vol. 40, no. 4, pp. 438–447, 2010.

[25] A. J. Greene, C. Prepscius, and W. B. Levy, "Primacy versus recency in a quantitative model: Activity is the critical distinction," *Learning and Memory*, vol. 7, no. 1, pp. 48–57, 2000.

[26] J. H. Herbert, "Data matters - specification and estimation of natural gas demand per customer in the northeastern United States," *Computational Statistics and Data Analysis*, vol. 5, no. 1, pp. 67–78, 1987.

[27] J. Hinman and E. Hickey, "Modeling and forecasting short-term electricity load using regression analysis," *Journal of Institute for Regulatory Policy Studies*, 2009.

[28] T. Hong, B. Liu, and P. Wang, "Electrical load forecasting with recency effect: A big data approach," working paper available online: www.drhongtao.com/articles, Tech. Rep., 2015, accessed: 2016-02-16.

[29] T. Hong, “Short term electric load forecasting,” Ph.D. dissertation, North Carolina State University, 2010.

[30] B. I. Ishola, R. J. Povinelli, G. F. Corliss, and R. H. Brown, “Identifying extreme cold events using phase space reconstruction,” *submitted to International Journal of Applied Pattern Recognition*, 2016.

[31] P. Kaefer, “Transforming analogous time series data to improve natural gas demand forecast accuracy,” Master’s thesis, Marquette University, Department of Mathematics, Statistics and Computer Science, Milwaukee, WI, May 2015.

[32] L. S. Kalkstein and K. M. Valimont, “An evaluation of summer discomfort in the United State using a relative climatological index,” *Bulletin of the American Meteorological Society*, vol. 67, no. 7, pp. 842–848, 1986.

[33] O. Kaynar, I. Yilmaz, and F. Demirkoparan, “Forecasting of natural gas consumption with neural network and neuro fuzzy system,” *Energy Education, Science and Technology*, vol. 26, no. 2, pp. 221–238, 2011.

[34] A. Khotanzad, H. Elragal, and T.-L. Lu, “Combination of artificial neural-network forecasters for prediction of natural gas consumption,” *Neural Networks, IEEE Transactions on*, vol. 11, no. 2, pp. 464–473, 2000.

[35] N. Liu, V. Babushkin, and A. Afshari, “Short-term forecasting of temperature driven electricity load using time series and neural network model,” *Journal of Clean Energy Technologies*, vol. 2, no. 4, 2014.

[36] Z. Liu and W. Zhu, “Homoclinic bifurcation and chaos in simple pendulum under bounded noise excitation,” *Chaos, Solitons & Fractals*, vol. 20, no. 3, pp. 593–607, 2004.

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

[38] S. Maitra and J. Yan, “Principle component analysis and partial least squares: Two dimension reduction techniques for regression,” *Applying Multivariate Statistical Models*, vol. 79, 2008.

[39] L. Marchi, M. Borga, E. Preciso, and E. Gaume, "Characterisation of selected extreme flash floods in Europe and implications for flood risk management," *Journal of Hydrology*, vol. 394, no. 1, pp. 118–133, 2010.

[40] S. E. Masten and K. J. Crocker, "Efficient adaptation in long-term contracts: Take-or-pay provisions for natural gas," *The American Economic Review*, vol. 75, no. 5, pp. 1083–1093, 1985.

[41] I. Matin, "Artificial neural network models to predict gas consumption," Master's thesis, Marquette University, Department of Electrical and Computer Engineering, Milwaukee, WI, November 1995.

[42] M. J. Mazerolle, "Appendix 1: Making sense out of Akaikes Information Criterion (AIC): Its use and interpretation in model selection and inference from ecological data," 2004, accessed: 2016-02-16. [Online]. Available: www.theses.ulaval.ca/2004/21842/apa.html

[43] Natural Gas Supply Association, "Natural Gas Distribution," 2013, accessed: 2016-02-16. [Online]. Available: http://naturalgas.org/naturalgas/distribution/

[44] ——, "Uses of Natural Gas," 2013, accessed: 2016-02-16. [Online]. Available: http://naturalgas.org/overview/uses/

[45] M. Ozturk, "Thermodynamic assessment of space heating in buildings via solar energy system," *Journal of Engineering and Technology*, vol. 1, no. 1, 2011.

[46] G. Palermo, P. Piraino, and H.-D. Zucht, "Performance of PLS regression coefficients in selecting variables for each response of a multivariate PLS for omics-type data," *Advances and Applications in Bioinformatics and Chemistry: AABC*, vol. 2, p. 57, 2009.

[47] B. Pang, "The impact of additional weather inputs on gas load forecasting," Master's thesis, Marquette University, Department of Electrical and Computer Engineering, 2012.

[48] A. D. Papalexopoulos and T. C. Hesterberg, "A regression-based approach to short-term system load forecasting," *Power Systems, IEEE Transactions on*, vol. 5, no. 4, pp. 1535–1547, 1990.

[49] R. J. Povinelli, “Identifying temporal patterns for characterization and prediction of financial time series events,” *Temporal, Spatial and Spatio-Temporal Data Mining: First International Workshop*, pp. 46–61, 2000.

[50] R. J. Povinelli and X. Feng, “Temporal pattern identification of time series data using pattern wavelets and genetic algorithms,” *Artificial Neural Networks in Engineering*, vol. 2, pp. 691–696, 1998.

[51] ——, “A new temporal pattern identification method for characterization and prediction of complex time series events,” *Knowledge and Data Engineering, IEEE Transactions on*, vol. 15, no. 2, pp. 339–352, 2003.

[52] R. J. Povinelli, M. T. Johnson, A. C. Lindgren, F. M. Roberts, and J. Ye, “Statistical models of reconstructed phase spaces for signal classification,” *IEEE Transactions on Signal Processing*, vol. 54, no. 6, pp. 2178–2186, June 2006.

[53] ——, “Statistical models of reconstructed phase spaces for signal classification,” *Signal Processing, IEEE Transactions on*, vol. 54, no. 6, pp. 2178–2186, 2006.

[54] R. J. Povinelli, M. T. Johnson, A. C. Lindgren, and J. Ye, “Time series classification using Gaussian mixture models of reconstructed phase spaces,” *Knowledge and Data Engineering, IEEE Transactions on*, vol. 16, no. 6, pp. 779–783, 2004.

[55] J. R. Quinlan, “Combining instance-based and model-based learning,” in *Proceedings of the Tenth International Conference on Machine Learning*, 1993, pp. 236–243.

[56] J. C. Robinson, “A topological delay embedding theorem for infinite-dimensional dynamical systems,” *Nonlinearity*, vol. 18, no. 5, p. 2135, 2005.

[57] T. Sauer, J. A. Yorke, and M. Casdagli, “Embedology,” *Journal of Statistical Physics*, vol. 65, no. 3-4, pp. 579–616, 1991.

[58] J. Sjöberg, H. Hjalmarsson, and L. Ljung, “Neural networks in system identification,” in *10th IFAC Symposium on System Identification, Copenhagen, Denmark, July, 1994*, vol. 2, 1994, pp. 49–72.

[59] M. Steinbach, L. Ertöz, and V. Kumar, “The challenges of clustering high dimensional data,” in *New Directions in Statistical Physics*. Springer, 2004, pp. 273–309.

[60] F. Takens, “Detecting strange attractors in turbulence dynamical systems and turbulence, Warwick 1980,” *Dynamical Systems and Turbulence*, vol. 898, pp. 366–381, 1981.

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

[62] E. C. Thom, “The discomfort index,” *Weatherwise*, vol. 12, no. 2, pp. 57–61, 1959. [Online]. Available: http://dx.doi.org/10.1080/00431672.1959.9926960

[63] R. D. Tobias, “An introduction to partial least squares regression,” in *Proc. Ann. SAS Users Group Int. Conf., 20th, Orlando, FL*. Citeseer, 1995, pp. 2–5.

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

[65] U.S. Energy Information Administration, “Space heating is the largest portion of household energy use,” 2014, accessed: 2016-02-16. [Online]. Available: http://www.eia.gov/todayinenergy/detail.cfm?id=18131

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

[67] J. Vondráček, E. Pelikán, O. Konár, J. Čermáková, K. Eben, M. Malỳ, and M. Brabec, “A statistical model for the estimation of natural gas consumption,” *Applied Energy*, vol. 85, no. 5, pp. 362–370, 2008.

[68] B. M. J. Will and J. C. Fransoo, “Operations management research methodologies using quantitative modeling,” *International Journal of Operations & Production Management*, vol. 22, no. 2, pp. 241–264, 2002.

[69] W. Zhang and X. Feng, “Predictive temporal patterns detection in multivariate dynamic data system,” in *Intelligent Control and Automation (WCICA), 2012 10th World Congress on.* IEEE, 2012, pp. 803–808.

[70] M. W. Zimmerman, R. J. Povinelli, M. T. Johnson, and K. M. Ropella, “A reconstructed phase space approach for distinguishing ischemic from non-ischemic ST changes using Holter ECG data,” in *Computers in Cardiology, 2003.* IEEE, 2003, pp. 243–246.