An Investigation into Water Consumption Data Using Parametric Probability Density Functions

by

Robert William Enouy

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Doctor of Philosophy

in

Earth Sciences

Waterloo, Ontario, Canada, 2018

EXAMINING COMMITTEE MEMBERSHIP

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner	Name:	Dr. Kent Novakowski
	Title:	Professor
Supervisor	Name:	Dr. Andre Unger
	Title:	Associate Professor
Co-supervisor	Name:	Dr. Rashid Rehan
	Title:	Associate Professor
Internal Member	Name:	Dr. David Rudolph
	Title:	Groundwater Chair & Professor
Internal-external Member	Name:	Dr. Mohammad Kohandel
	Title:	Associate Chair of Undergraduate Studies & Associate Professor
Other Member(s)	Name:	Dr. Marios Ioannidis
	Title:	Associate Chair of Undergraduate Studies & Professor

AUTHOR DECLARATION

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

ABSTRACT

Many datasets can be expressed through the usage of frequency-based histograms that provide compact visualizations of large volumes of data. A motivating hypothesis of this work is that the relative frequency of measurements represents important information for characterizing causal relationships. The basis of this work is transforming discrete histograms into continuous probability density functions (PDFs) using conservation of probability. In essence, this thesis will investigate whether conservation of probability can be applied as a governing law that characterizes how both physical and abstract measurement histograms will evolve through time.

The main application within this thesis involves transforming bimonthly residential water consumption histograms into parametric PDFs for 60 sequential billing periods. Consistent parameterization for each billing period allows for regression analysis to infer a causal relationship between the PDF statistics and ambient conditions such as price and weather. This method generates partial differential equations (PDEs) for each statistic that combine to reproduce measurement data PDFs for varying ambient conditions through time using an "advection-dispersion" like relationship. The significance of this methodology is that parametric PDEs can describe the historical relationship between the measurement PDF and ambient conditions. This relationship may be exploited in future work to generate parametric PDEs that forecast the evolution of measurement PDFs through their location, scale, and shape with respect to influential ambient conditions.

This thesis also demonstrates a relationship between measurement PDFs and the governing PDEs for the physical process of molecular diffusion. This outcome provides compelling evidence that conservation of probability applies to both abstract and physical systems, which suggests conservation of information is a unifying concept for modeling systemic response. Ultimately, conservation of probability provides a mechanism for reconciliation that ensures no information is either created or destroyed, while generating PDEs to reproduce measurement data as spatially- and temporally-continuous PDFs.

ACKNOWLEDGEMENTS

I would like to express my appreciation to Dr. Andre Unger and Dr. Rashid Rehan for their valuable and constructive suggestions during the planning and development of this research.

I would also like to thank the City of Waterloo and the water utility staff for the generous donation of water consumption data and insight into residential consumer preferences. This research project would not have been possible without their contributions.

The precipitation and temperature data used in this thesis were obtained in parts from the NASA Langley Research Center POWER Project funded through the NASA Earth Science Directorate Applied Science Program; the University of Waterloo Weather Station; and Environment and Climate Change Canada for Kitchener/Waterloo Weather Station, Ontario, Canada.

TABLE OF CONTENTS

LIST OF FIGURES VII

LIST OF TABLES VIII

1. INTRODUCTION 1

2. TRANSFORMING HISTOGRAM DATA INTO PARAMETRIC PDFS 8

2.1. THEORY 12

2.1.1. THE MEDIAN AND STANDARD DEVIATION 12

2.1.2. CONTINUOUS DIFFERENTIABLE PDFS 13

2.1.3. STATISTICAL TRANSFORMATIONS 15

2.1.4. THE CONTROL FUNCTION 17

2.1.4.1. The Normal Distribution 18

2.1.4.2. α_2 Family of Curves 19

2.1.4.3. Polynomial Series Extension 21

2.1.4.4. Modified-Fourier Series Extension 22

2.1.5. MEDIAN-RELATIVE SPACE 22

2.1.5.1. Data Culling 23

2.1.5.2. Objective Function 24

2.1.5.3. The Mean Statistic 24

2.1.6. DEGREES OF FREEDOM ANALYSIS 26

2.2. APPLICATION 28

2.3. CONCLUSIONS 38

3. ADVECTIVE-DISPERSIVE TRANSPORT OF A PROBABILITY DENSITY FUNCTION: MODEL DEVELOPMENT AND WATER CONSUMPTION APPLICATION 40

3.1. MODEL DEVELOPMENT 44

3.1.1. DISCRETE DATA STATISTICS 46

3.1.2. TIME-CONTINUOUS STATISTICS 47

3.1.3. ADVECTIVE-DISPERSIVE TRANSPORT WITH AMBIENT PROCESSES 49

3.2. RESIDENTIAL WATER CONSUMPTION APPLICATION 51

3.2.1. PARAMETRIC PDFS AS A REPRESENTATION OF THE CONTINUUM RESPONSE 52

3.2.2. PRICE AND WEATHER AS AMBIENT PROCESSES 57
3.2.2. ADVECTIVE-DISPERSIVE TRANSPORT MODEL PARAMETERIZATION 59
3.3. DISCUSSION 64
3.4. CONCLUSIONS 70

4. CONSERVATION OF PROBABILITY AND PARAMETRIC PDES 72

4.1. THEORY 74
4.1.1. NORMALLY DISTRIBUTED MEASUREMENTS 77
4.1.2. SCALING THE BROWNIAN MOTION DIFFUSION COEFFICIENT 79
4.1.2.1. One-dimensional Diffusion 80
4.1.2.2. Radial Diffusion 82
4.1.2.3. Probabilistic Area-based Diffusion 84
4.1.3. FICK'S LAW 89
4.3. DISCUSSION 92
4.4. CONCLUSIONS 94

5. THESIS CONCLUSIONS 95

REFERENCES 97

APPENDIX A 102

APPENDIX B 108

APPENDIX C 122

LIST OF FIGURES

Figure 1.1: Schematic visualization of data management approach 4
Figure 1.2: Mean bimonthly residential water consumption for the City of Waterloo 2007-2016 5

Figure 2.1: Histogram data for disparate datasets 10
Figure 2.2: Raw data PMF and parametric PDF for disparate datasets 11
Figure 2.3: Bounding probability density functions for the $\boldsymbol{\alpha_2}$ family of curves 20
Figure 2.4: Corresponding control functions for bounding PDFs of the $\boldsymbol{\alpha_2}$ family of curves 21
Figure 2.5: Control function visualization for each histogram 33
Figure 2.6: Standard-score PDF visualization for each histogram 33
Figure 2.7: Visualization of the CMF and CDF for the optimally parameterized PDFs 36

Figure 3.1: Select residential water consumption histograms for November/December from the City of Waterloo, Ontario, Canada 42
Figure 3.2: Select water consumption data and optimal parametric fit 54
Figure 3.3: Select optimally parameterized PDFs for November/December billing period 55
Figure 3.4: Select optimally parameterized PDFs showing seasonal influence on water consumption 56
Figure 3.5: Ambient conditions for water consumption during entire analysis period 59
Figure 3.6: Curvilinear model results for the median and standard deviation 65
Figure 3.7: Transport model results for select November/December periods 66
Figure 3.8: Transport model results for select July/August periods 67
Figure 3.9: Transport model results for mean water consumption 68
Figure 3.10: Transport model results for 2015-2016 May/June and July/August periods 69

Figure 4.1: Water consumption data and parametric PDF for July/August 2007 75
Figure 4.2: A standard normal random walk in two-dimensions 82
Figure 4.3: Geometric representation of area-based PDF 86

LIST OF TABLES

Table 2.1: Data transformations between each spatial orientation 16
Table 2.2: Data boundaries in each spatial orientation 17
Table 2.3: Parametrization of the normal distribution and $\boldsymbol{\alpha_2}$ family of curves 19
Table 2.4: Degrees of freedom analysis 27
Table 2.5: Summary of statistics for the four disparate datasets 29
Table 2.6: Exponential polynomial and modified-Fourier series control functions for each application 32
Table 2.7: Parameterization for the water demand, hydraulic conductivity, and S&P 500 datasets 34
Table 2.8: Parameterization for the Lenna light intensity dataset 35

Table 3.1: MSE values between observed and optimally parameterized mean statistic 53
Table 3.2: Summary of model parameterization for Equation 3.12 61
Table 3.3: Active model parameters for each statistic 62

Table 4.1: Second-order Homogeneous PDEs for various physical processes 79
Table 4.2: Scaled diffusion coefficients and PDEs for various dimensionality 91

1. INTRODUCTION

Advancements in water distribution have been a prerequisite for sociocultural evolution throughout history; from nomadic tribes, through irrigation and farming, to waterside villages and aqueduct-supplied cities, and finally piped home delivery. In general, these advances have consistently improved health, standard of living, and life expectancy by reducing exposure to disease and promoting hygiene. As a result of continual investment into these systems, potable water has never been as widely accessible as it is today. The need to sustain and extend modern potable water distribution systems should make the development of appropriate management practices an important priority. The sustainability of water distribution systems requires substantial investments into the maintenance and capital replacement of infrastructure. Water distribution systems have vastly expanded around the world into undeveloped areas, often with little forethought to the future necessity for maintenance of the system. For instance, a large proportion of water infrastructure developed in the post-World War II era is currently nearing the end of its service life and requires replacement (AWWA, 2012). The American Water Works Association report "Buried No Longer" estimated that at least $1 trillion dollars for maintenance and replacement would be required over the next 25 years in the United States to maintain current levels of water service. The increased costs of operating and maintaining these systems are expected to be passed onto the consumer through a "pay-as-you-go" approach whereby water prices will be increased to generate higher revenues with the expectation of matching these system expenses.

Revenues within municipal utilities are subject to variability that can result from a number of factors: changing population, consumer demand changes, and water price structures (Eskaf et al. 2014). Historically speaking, government policy levers such as the US Energy Policy Acts of 1992 and 2005 and the US Energy Independence and Security Act of 2007 have been key drivers in reducing residential and commercial water consumption (Hunter et al. 2011). Beecher et al. (2012) state that flat or declining sales are affecting many utilities and the loss of revenues caught many utility managers and industry analysts off guard. With declining revenues, utilities have little choice but to raise their water rates or implement alternative water pricing structures. Pricing structures such as increasing block pricing (IBP), decreasing block pricing (DBP), and volume

constant pricing structures (fixed and variable components) are applied throughout North America. These pricing structures have been investigated to understand their influence on revenue sustainability and water affordability (Mayer et al., 2008; Beecher et al., 2010; Beecher et al., 2012; Mehan III et al., 2012; Eskaf et al,, 2014).

When water utilities are forced to increase their water rates beyond cost-of-living inflation, aggregate water consumption generally decreases. Under these conditions, consumers switch to low-flow appliances and limit unnecessary water use to reduce their water bills. Financial forecasts that fail to consider this effect on water demand will therefore overstate anticipated system revenues and potentially lead to realized shortfalls. In standard microeconomic theory, price elasticity of demand is used to describe the relationship between price and demand. Previous studies have attempted to provide insights into water distribution and wastewater services including the development of system dynamics models (Rehan et al., 2011; 2013; 2014a; 2014b; 2014c), and studies on the price elasticity of domestic water demand (Boland et al., 1984; Espey et al., 1997; Dalhuisen et al., 2003; Brookeshire et al., 2002; Olmstead et al., 2007; Worthington et al., 2008; Serbi, 2014). These efforts have significantly contributed to understanding the intricacies of asset management and water price setting. Moreover, the need for utilities to balance revenues with expenditures has motived the systematic recording of water consumption data by residential, commercial and institutional account holders. This allows the utility to infer how utility-wide water consumption is responding to factors such as real price increases in the unit cost of water.

Advancements in data collection techniques through automation and innovative technologies provide researchers with a wealth of information to develop, infer, and confirm theoretical inquiry. Many datasets can be expressed through the usage of frequency-based histograms that provide compact visualizations of large volumes of data. Some examples may include the residential income distribution, residential water consumption measurements, and stock market capitalization indices, among many more. Current data analysis techniques rely heavily on linear regression and the normal distribution to show correlation and infer causality between otherwise unconnected measurements. However, dataset histograms often do not conform to the normal distribution and current methodologies may be insufficient for explaining some observations and intuitive relationships. A motivating hypothesis of this work is that the relative frequency of measurements represents important information for characterizing causal

relationships. In this context, the normal distribution represents a specific shape, or relative frequency, that may not reflect real-world observations.

The motivation for this thesis is to develop a methodology for parameterizing probability density functions (PDFs) that represent histograms of utility-wide residential water consumption, where these histograms exhibit asymmetry and heavy-tail shape attributes. These parametric PDFs include but also extend beyond the normal distribution in order to accurately reproduce measurement data histograms. This process involves measuring the median as an indicator of location, the standard deviation as a measure of scale, and a series of parameters that quantify the shape of the PDF. Furthermore, it is the intent of this thesis to explore the possibility that PDFs representing real-world measurements are the solution to an "advective-dispersive" like process that continually evolve through time. This exploration proceeds by fitting experimental data with parametric PDFs and tracking the influence of ambient conditions on the statistics that describe the PDF. For instance, the influence of price and weather conditions on the location, scale and shape of the residential water consumption PDF. A hypothesis of this work is that the evolution of each statistic can be described by a partial differential equation (PDE), where the PDEs model the historical relationship between the statistic as a function of ambient conditions and time. Furthermore, these statistics and their representative PDE can then be combined to produce parametric PDFs that accurately model the probability of observing a prescribed magnitude of measurement data over anticipated ambient conditions at future dates. In the context of the water consumption problem, this relationship can be exploited to forecast the evolution of the water consumption PDF with respect to anticipated ambient water price and weather conditions.

An outcome of this work is a series of software programs to execute the theory developed herein by reproducing discrete datasets as continuous PDFs and performing multivariate regression to understand the functional relationship between the statistics that combine to produce each PDF with respect to ambient conditions and time. This theory was initially tested using Excel® and validated with independent development in Matlab®. The complexity of the problem and sheer magnitude of water consumption data has motivated implementation of this theory using SQL database management software and the object-oriented Python environment. This thesis focuses on the theory behind the software development and only provides superficial details about software development. Each software approach produces the same analysis results within some margin of error as a consequence of using slightly different optimization algorithms, which are all ultimately

based on a mean-squared error approach. Notably, the Python implementation provides the most general algorithmic design, which has been tested using additional water consumption data, as well as with other applications including daily S&P500 data, daily NASDAQ data, patient health records, image analysis, and hydraulic permeability data. Figure 1.1 presents a schematic visualization for automation of data management and regression analysis.

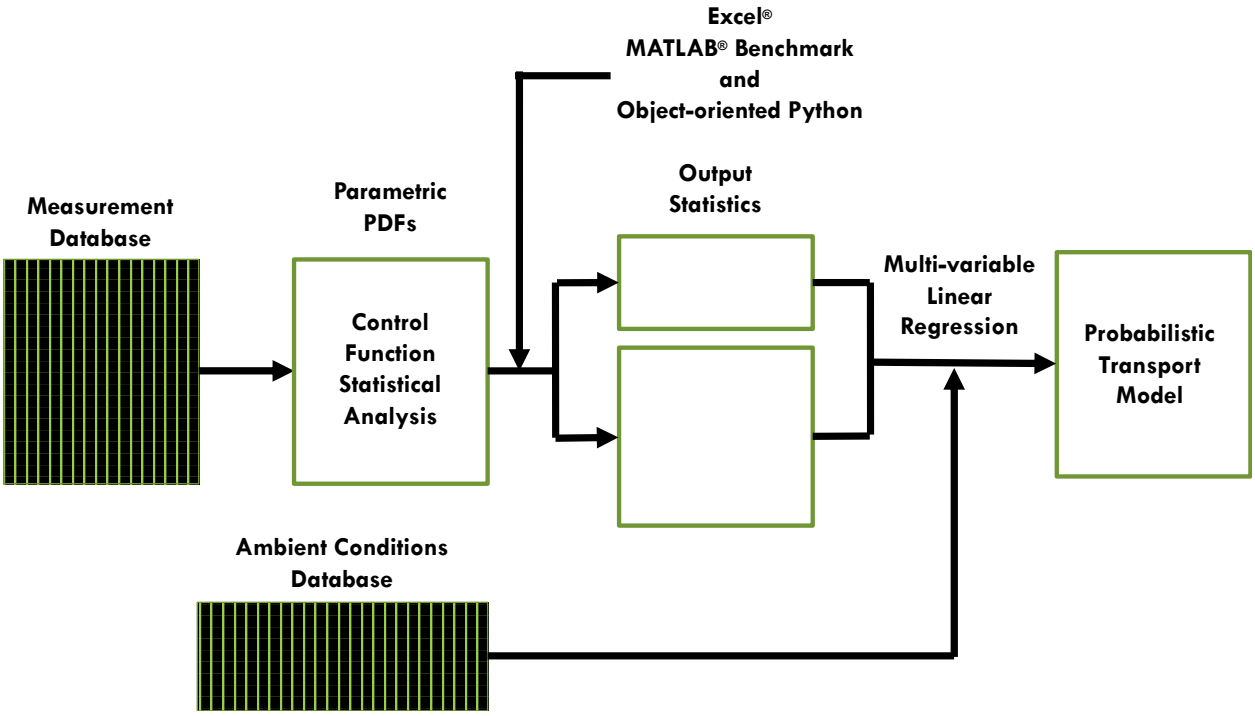

Figure 1.1 – Schematic visualization of data management approach.

Although the algorithm developed is generally applicable, the motivation of this work is realized by specifically focusing on how residential water consumers respond to changes in price and weather. Forecasting models for water consumption that consider the influence of price and weather are important for helping water utilities to develop management strategies that promote financial and resource sustainability. The water utility within the City of Waterloo, Ontario, Canada provided this research opportunity with residential water consumption specified for individual accounts within the utility spanning a period of 10 years. The data is collected on bimonthly intervals, which provides a natural temporal grouping of measurements for this analysis. Combining bimonthly observations into frequency histograms reveals that the consumption datasets reflect asymmetric and shifted distributions with heavy tails. Each bimonthly period

represents a unimodal distribution that continuously changes its location, scale and shape as the ambient conditions of price and weather evolve through time. Notably, the mean statistic of water consumption can be estimated directly from these distributions; therefore, understanding how the distributions change may provide insight into the evolution of their mean statistic. Intuition may indicate that both price and weather should have a significant influence on how the water consumption distribution evolves and this expectation is supported when visualizing a time-series representation of the mean water consumption statistic. Figure 1.2 shows that the arithmetic mean water consumption statistic exhibits a wave-like response, similar to that of seasonal weather variations, superimposed upon a continual decline in water consumption that is presumed to correspond with annual water price increases.

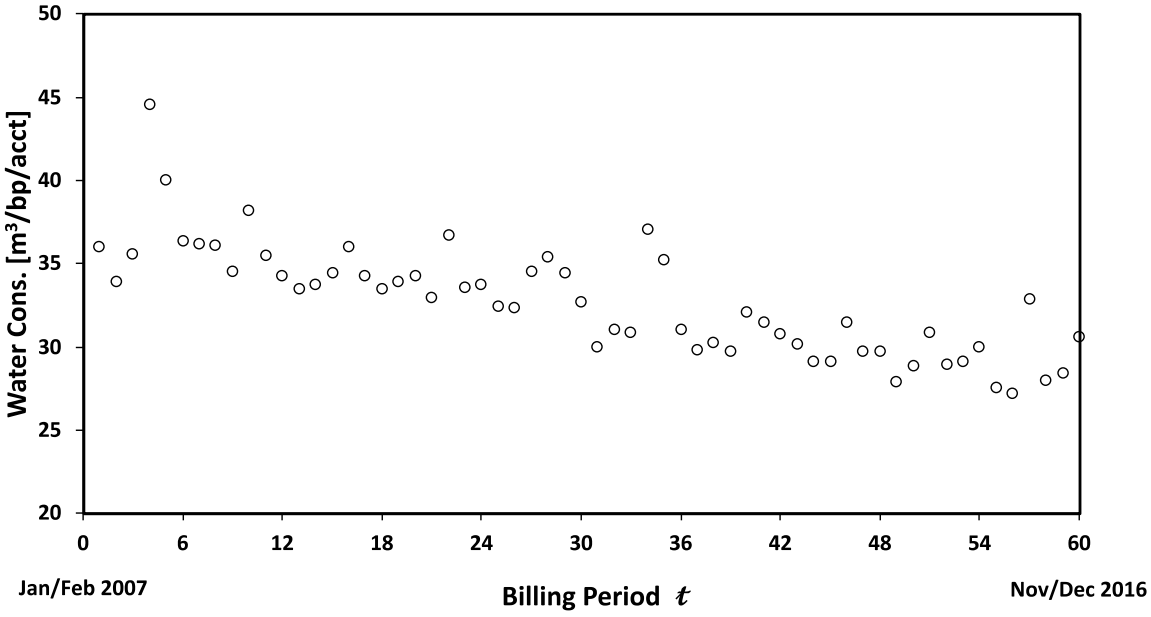

Figure 1.2 – Mean bimonthly residential water consumption for the City of Waterloo 2007-2016.

The progression of this thesis is organized and articulated in the following three chapters. Chapter 2 focuses on transforming histogram data into a parametric PDF. Chapter 3 then follows with the development of an "advective-dispersive" like transport model describing residential water consumption. Finally, Chapter 4 involves a discussion on the relationship between conservation of probability and the resulting parametric PDEs.

The contributions of the second chapter largely focus on parameterizing a shape-controlling function, through conservation of probability, that allows fitting histogram data as a continuously differentiable function. This methodology is generally consistent with standard kernel density estimation techniques. However, a hierarchical relationship between the shape-controlling function, PDF, and its cumulative distribution function (CDF) are relied upon to generate the density estimate. An ordinary differential equation (ODE) transforms the shape-controlling function into the PDF. A subsidiary but important contribution of this chapter is the development of the median-relative space to automate data culling and the parameterization of the shape-defining function for disparate datasets as shown in Figure 1.1. Finally, the intent of the second chapter is to demonstrate that the PDF and its associated mean statistic are solely a function of the median (location), scale (standard deviation), and shape-controlling function. As a result of this outcome, this thesis views both the PDF and mean statistic through the lens of an "advective-dispersive" like process. Notably, changes to the median reflect bulk advection, whereas changes to the standard deviation and shape-controlling function constitute dispersion.

The contributions of the third chapter follow the motivation of this thesis and demonstrate that the water consumption histogram data for all sampling time intervals can be transformed into PDFs using a consistently parameterized shape function. These PDFs exhibit asymmetry and heavy-tail shape attributes and hence do not conform to a normal distribution. Note that the parametric PDFs are derived from the entire population of residential consumers. Hence, the shape controlling functions capture important information representing the PDF, its mean statistic, and total water consumption trends. An outcome is that the parametric PDF changes location, scale and shape under the influence of temporal variability in the ambient conditions of price and weather. Performing curvilinear regression on the median, standard deviation, and shape parameters provides a statistically significant correlation with price and weather variables. Recombing the resulting regression relationships for each statistic into a transport model accurately reproduces the measurement histogram for most billing periods. Therefore, an outcome of this chapter is the development and parametrization of a regression function that describes the changing shape of the water consumption PDF. Specifically, multi-variate curvilinear regression is used to parameterize PDEs for each statistic as a function of ambient price and weather information. Therefore, this analysis provides empirical evidence of a parametric PDE-PDF relationship that may provide insight into other physical or abstract applications.

The fourth chapter investigates the relationship between a scaled normally distributed PDF as the solution to second-order homogeneous PDEs. This relationship is based on three key ideas: 1) conservation of probability in a measurement space; 2) spatial-continuity of measurement data; and, 3) temporal-continuity of measurement data. The objective of this methodology is to build upon the current knowledge related to Brownian motion as being a representation of a second-order homogeneous PDEs, thereby reinterpreting the experimental water consumption histogram data from Chapter 3 to being the solution to a probabilistic transport process. This chapter uses the Einsteinian diffusion PDE as the basis for investigating molecular diffusion relative to artificially-generated molecular displacement data, under the premise that the measurement data conforms to a normal distribution. The analysis shows that the solution to the Fourier heat transfer equation in solids may also describe molecular diffusion in two dimensions. Furthermore, an extension of this solution is adapted to reproduce Fick's Law as a result observing the probabilistic process of Einsteinian diffusion in two dimensions. The outcome of the fourth chapter is to demonstrate that the approach from Chapter 3, which generates a parametric PDE for an abstract economic system, yields coefficients that seem tangible in much the same manner as the diffusion coefficient in the context of the Einsteinian diffusion PDE.

The outcome of this thesis provides compelling evidence that conservation of information is a unifying concept for modeling system response. This thesis concludes with a discussion on how PDFs the constitute the solution to PDEs of abstract and physical processes that reflect the interaction between three system components: 1) a source/sink term, 2) the measured property, and 3) a conduit that connects the source/sink to the measured property. Therefore, a conclusion is that the parameters within the governing PDE meaningfully describe and quantify the properties of the conduit. For the water consumption application, the conduit is the household-specific qualities that compel water consumption, the source term is the necessity to consume water to maintain standard of living conditions, and the measurement is the volume of water consumed in a billing period. Future work could exploit parameterization of the PDE to forecast pairs of source/sink terms and the resulting solution of a parametric PDF. In this context, evaluating parametric PDFs as a solution to PDEs may be a more general expression of systemic response than what could be achieved by constraining systemic response to be controlled by second-order homogeneous PDEs.

2. Transforming Histogram Data into Parametric PDFs

Collectively, individual measurements obtained within a discrete sampling interval define the range of system conditions. In the context of this work, these measurements are non-zero and real-valued observations and are subject to measurement error. Subsequently, evaluating the ordered frequency of these measurements constructs a histogram. Dividing the frequency at which measurements occur within a discrete sampling interval by the total number of measurements transforms the histogram into a probability mass function (PMF). The probability of observing a measurement within a range of discrete intervals is realized by summation, which results in the corresponding cumulative mass function (CMF). The utility of parametric probability density functions (PDFs) is that they provide an empirical mechanism to mathematically characterize the defining attributes of discrete datasets. These attributes include the location, scale, and shape of the histogram, which translate into statistics that combine to accurately express PMFs as continuous PDFs.

Kernel density estimation (KDE) is a non-parametric approach to smoothly translate random sample data into a continuous PDF using additive normal distributions with smoothing parameters. Zambom and Dias (2012) provide a thorough review of the KDE methodology in the context of econometrics. KDE is sufficiently versatile to accurately reproduce many empirical datasets characterized by asymmetric, tail-heavy, or multi-modal PMFs. However, the KDE methodology does not provide a consistent parametrization when the measurement PMF changes its location, scale and shape in response to transient ambient conditions. Hence, it is impossible to regress this inconsistent parameterization and infer the functional relationship of casual influences that drive the response of the PMF. Attempts to address this shortcoming include applying modified conditional density estimators (Hyndman et al., 1996) and multivariate function estimation (Stone, 1994). The goal of Stone was to extend generalized linear modeling to handle multivariate data involving response variables and covariates that include a mixture of continuous and categorical variables. The intent is to consider the variables and covariates that reproduce the fitted PMF as an empirical treatment of some probabilistic process that evolves through time under the influence of ambient processes. Duda et al., (2017) introduce a methodology for fitting parametric PDFs to characterize asymmetric, tail-heavy, and multi-modal PMFs. Their

methodology uses polynomial and Fourier series that are multiplied by Gaussian distributions. However, the issue of consistency in parametrization remains.

This work develops a general methodology to reproduce histogram data that are asymmetric, shifted, tail-weighted, or even multi-modal as parametric PDFs. The main contribution of this methodology relative to the existing body of non-parametric and parametric PDF estimators is that a single parametric function can fit many time-sequential experimental data sets representing the temporal evolution of a single probabilistic process. This addresses the consistency of parameterization issue identified above. The objective of this chapter is to introduce the methodology underpinning this framework which is based on conventional statistics and calculus, with its versatility illustrated by application to four disparate data sets from economics, engineering, finance, and image analysis. The histograms shown on Figure 2.1 comprise datasets that range from hundreds to thousands of measurements denoted as x_i, and vary in complexity from unimodal to multi-modal distributions. The discrete intervals on each histogram represent the probability of occurrence within the PMF when the histogram frequency is divided by the total number of measurements N_i. Equation 2.1 defines discrete intervals within a histogram and shows how these discrete bins relate to the PMF $p_{x,k}$ and CMF c_{x,k_1}.

$$
\begin{aligned}
&h_{x_{k-1}<x_i<x_k} \equiv \text{frequency within histogram bin} \\
&\text{PMF,} \qquad p_{x,k} = \frac{h_{x_{k-1}<x_i<x_k}}{N_i}, \qquad 0 \le p_{x,k} \\
&\text{CMF,} \qquad c_{x,k_1} = \sum_{k=1}^{k_1} p_{x,k}\,, \qquad 0 \le c_{x,k_1} \le 1,
\end{aligned}
\tag{2.1}
$$

where $h_{x,k}$ represents the frequency of measurement values x_i within the discrete sampling interval $x_{k-1} < x_i < x_k$; the PMF $p_{x,k}$ divides each histogram bin by the number of observations N_i; and, the CMF c_{x,k_1} follows by summing over the bins from $k = 1 \to k_1$. In fact, Figure 2.2 presents the associated PMF of each dataset. Notably, these representations also include the eventual outcome of this analysis, which is a parametric PDF that accurately reproduces each dataset as a continuously differentiable function that implies data compression. Here, data compression refers to the ability to reproduce the discrete dataset as a continuous function that contains the same

information as the measurement data, while using less parameters than number of measurement data.

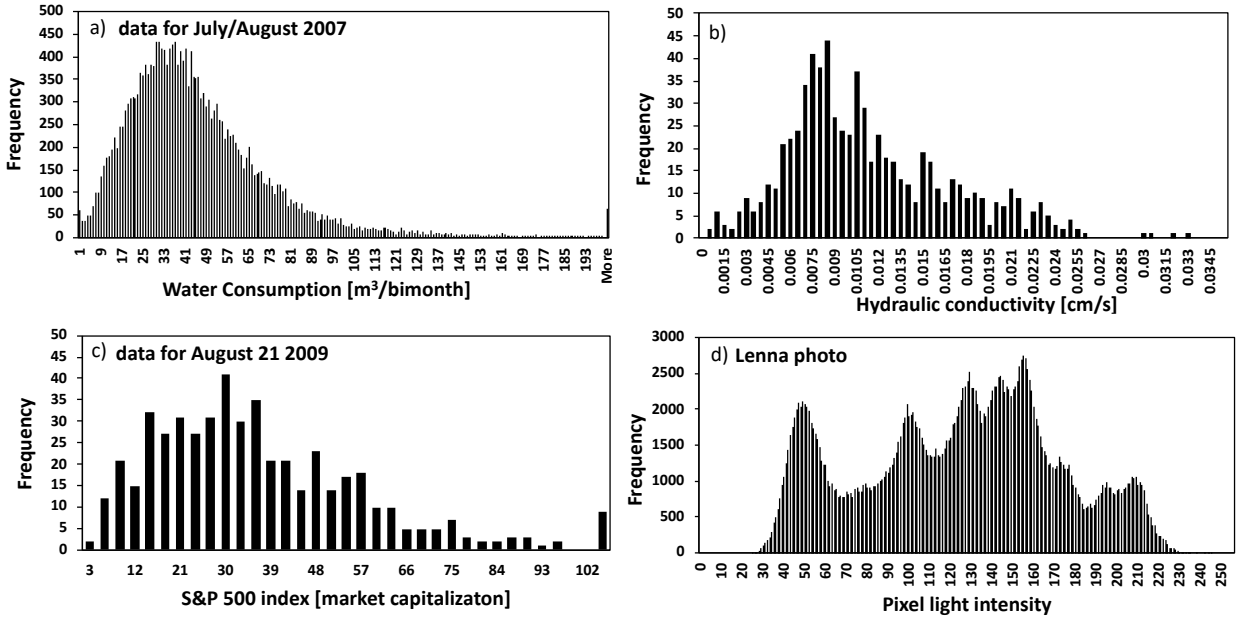

Figure 2.1: Histogram data for disparate datasets. Figure 2.1.a depicts single-family residential water consumption data from the July/August bimonthly billing period within the City of Waterloo, Ontario, Canada; Figure 2.1.b presents hydraulic conductivity measurements obtained from section cores drilled along a single cross-section within the Borden aquifer (Sudicky, 1986); Figure 2.1.c shows S&P 500 market capitalization index values obtained from information collected by Stockwiz (2009) on August 21, 2009; and, Figure 2.1.d illustrates light intensity data obtained from the classic "Lenna" photograph (Hutchinson, 2001).

The parametric control function represents the foundation of the proposed framework because it generates continuously differentiable PDFs in the standard-score space (see Figure 2.2). The control function embodies parametrization that replicates the shape of the PMF and CMF; and, hence the probability of occurrence within any interval on the histogram. The relationship between the control function and PDF is specified by an ordinary differential equation (ODE), where the control function is the lognormal derivative of a PDF with respect to the standard-score variable z. This relationship defines how the shape of the distribution will change along the standard-score axis. The utility of the control function is that it defines the shape attribute and provides an empirical mechanism to reproduce discrete datasets as continuous functions. The

median and standard deviation project the standard-score z PDF into the measurement x and median-relative y spatial orientations. Together, the median, standard deviation, and control function provide sufficient information to specify the hierarchical relationship between the control function, PDF, and CDF simultaneously in all spatial orientations; x, y and z. The proposed framework provides efficient parametric compression of discrete datasets without assuming a predefined distribution shape. This ultimately reduces the possibility of information loss associated with statistics that describe non-Gaussian datasets, while simultaneously reducing the storage needs to maintain data fidelity.

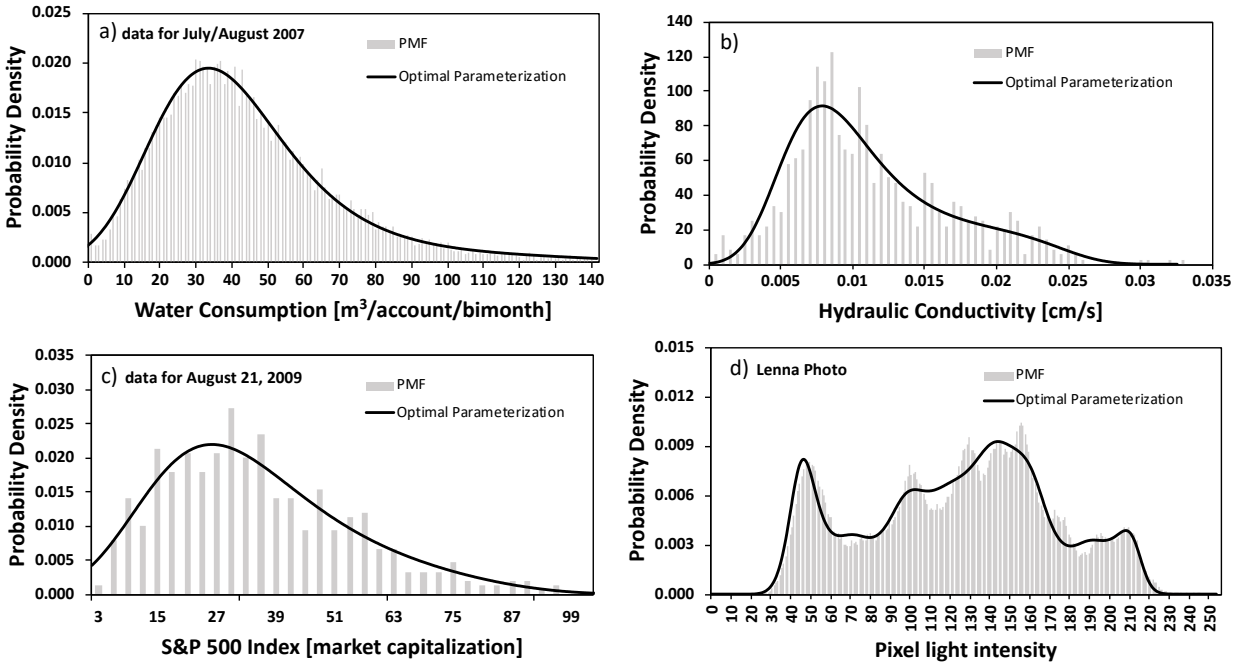

Figure 2.2: Raw data PMF and parametric PDF for disparate datasets. Figure 2.1.a depicts single-family residential water consumption data; Figure 2.1.b presents hydraulic conductivity measurements; Figure 2.1.c shows S&P 500 market capitalization index values; and, Figure 2.1.d illustrates light intensity data.

The outline for this work begins with the theoretical development of the framework. The utility of the framework is empirically demonstrated through its application to the PMFs shown on Figures 2.1 and 2.2.

2.1. THEORY

This work combines statistics and calculus to reproduce discrete histogram data as a continuously differentiable parametric PDF. This section reimagines ubiquitous relationships in statistics and develops a framework for evaluating the shape of a discrete dataset as a mathematical function. Attributes of this work pertaining to statistics and calculus are organized into separate sections below. The outcome from this theory is fourfold. First, this work introduces the control function, which characterizes the slope of a continuously differentiable PDF in the standard-score space. Second, PDFs are wholly defined by their representative statistics: the median, standard deviation, and control function, which are measures of location, scale, and shape, respectively. Third, the hierarchical integral relationship between the control function, PDF, and CDF allows this theory to compress the information embodied by the discrete histogram data into minimal sets of information. Fourth, the mean value is entirely dependent upon the combination of the above statistics, which provides the basis for developing causative models that do not rely upon Gaussian distributions. This section progressively addresses these key outcomes.

2.1.1. THE MEDIAN AND STANDARD DEVIATION STATISTICS

The median is arguably the simplest statistic associated with a discrete dataset, and is a measure of its location or central tendency with no assumption regarding its shape. Equation 2.2 introduces a heuristic to evaluate the median $m_{x,i}$ of a discrete dataset as:

$$m_{x,i} = \left\{\frac{N_i + 1}{2}\right\}^{th} \text{ value} \tag{2.2}$$

Where, N_i is the number of discrete measurements "i" within the dataset. If a dataset has an even number of discrete measurements, the median will be the average of the two middle data points.

The standard deviation is also a simple statistic associated with a discrete dataset, and defines its scale, also with no assumption about its shape. Equation 2.3 presents a modified version of the standard deviation $\sigma_{x,i}$ of a discrete dataset about its median value $m_{x,i}$ as:

$$\sigma_{x,i} = \sqrt{\frac{1}{(N_i - 1)} \sum_{i=1}^{N_i} \left[x_i - m_{x,i}\right]^2} \tag{2.3}$$

Where, x_i is the magnitude of measurement "i" obtained in the measurement and dimensional x space. Pearson (1894) first introduced the standard deviation of the dataset x_i relative to the arithmetic mean $\mu_{x,i}$. Equation 2.3 combines the median absolute deviation introduced by Gauss (1816) with the idea of squaring the deviation from Pearson (1894). Equation 2.3 is equivalent to Pearson's interpretation of standard deviation for symmetric distributions where $m_{x,i} = \mu_{x,i}$. The analysis proceeds with the above modification of the standard deviation being relative to the median given that both $m_{x,i}$ and $\sigma_{x,i}$ operate on the discrete elements of the dataset x_i, while $\mu_{x,i}$ measures the scalar continuum condition of the system. Later, this chapter clearly demonstrate that the mean statistic is a function of the median, standard deviation, and control function. In this spirit, evaluating the standard deviation as a function of the median prevents a recursive relationship between the standard deviation and mean value for asymmetric datasets.

The subsequent section on statistical transformations further illustrates the importance of the median and standard deviation. Specifically, the standard deviation transforms PMFs and PDFs between the measurement space x and standard-score space z. Furthermore, this chapter introduces a new transformation, herein referred to as the median-relative space y, which normalizes PMFs and PDFs by dividing each measurement/position by the median statistic to produce a dimensionless dataset. While both y and z are non-dimensional representations of x, they have different implications in relating PMFs to PDFs. The following sections discuss the implications and merits of defining the shape of the PDF in the standard-score space through the control function.

2.1.2. CONTINUOUSLY DIFFERENTIABLE PDFS

Earlier, this analysis defines PMF intervals $p_{x,k}$ to represent the probability of finding a discrete measurement x_i within the k^{th} bin of the histogram. This section introduces the ideas and notation that transform PMFs $p_{x,k}$ into continuously differentiable PDFs p_x over the full range of the measurement space. To achieve this goal, this equivalence must first be expressed in the standard-score space as $p_{z,k}$ which involves multiplication of $p_{x,k}$ by $\sigma_{x,i}$: $p_{z,k} = \sigma_{x,i}\, p_{x,k}$.

Evaluating the equivalence between the PMF and PDF $p_{z,k} \cong p_z$ in the standard-score space z is advantageous because its central-tendency is zero and it is therefore conducive to reproducing symmetry – for example, the normal distribution. This chapter contends that the standard-score space is the appropriate spatial reference for parametrizing the control function and the resulting PDF. Given its ubiquitous nature, the standard-score space z is used here before it is explicitly defined in the following section.

To begin, this section introduces a parametric control function g_z that mathematically characterizes the slope of the PDF p_z in the standard-score space z. The control function g_z is simply a component of an ODE that is consistent with the derivative of Gauss' maximum likelihood estimator for the error process (Gauss, 1809) and Stahl's derivation of the normal distribution (Stahl, 2006). Equation 2.4 represents the relationship between the lognormal derivative of the PDF $\frac{1}{p_z}\frac{dp_z}{dz} = g_z$ and its corresponding indefinite integral with respect to the standard score variable z. Appendix A.1 presents a step-by-step derivation from the concept of the control function g_z to the general form of the PDF $p_z = \exp(\int g_z\, dz)$ as:

$$\frac{dp_z}{dz} = g_z p_z \implies p_z = \exp\left(\int g_z\, dz\right) \tag{2.4}$$

The indefinite integral in Equation 2.4 solely serves the purpose of providing a consistent transformation between control function parameterization and the corresponding PDF p_z. While this integration has no influence on convergence of the PDF to unit area, it provides a functional form that is able to match the shape considerations of the empirical data being reproduced as a continuous function. Convergence of this relationship to unit area relies upon integration of the PDF p_z on a definite interval within the standard-score space z. This process then adjusts the constant of integration from evaluating $\int g_z\, dz$ to scale the definite integral to unit area. Equation 2.5 introduces the relationship between the PDF p_z and its corresponding CDF c_z as:

$$p_z = \exp\left(\int g_z\, dz\right) \implies c_z = \int_{z_0}^{z_1} p_z\, dz \tag{2.5}$$

where, the integration is defined on the interval of $z_0 \leq z \leq z_1$, and z_0 represents the standard-score position pertaining to the origin of the discrete data in the measurement space x_0. The

relationship between control function g_z, PDF p_z and CDF c_z projects into the measurement space yielding their measurement space equivalent PDF p_x and CDF c_x using the standard deviation $\sigma_{x,i}$. At this point, this analysis has generally qualified the attributes relating to the location, scale, and shape of PDFs through the median, standard deviation, and control function, respectively. The following sections fully define the statistical transformations between the measurement space x, median-relative space y, and standard-score space z, and expresses the control function as a parametric relationship that easily produces asymmetric, shifted, tail-weighted, and even multi-modal distributions.

2.1.3. STATISTICAL TRANSFORMATIONS

The median and standard deviation statistics transform PDFs between the measurement space x, the median relative space y, and the standard-score space z. A key attribute of this relation is that the CDF is identical in each spatial representation, which ensures conservation of probability of occurrence for all spatial representations as:

$$\int {p_x}^* \, dx = \int {p_y}^* \, dy = \int p_z \, dz \tag{2.6}$$

Where, ${p_x}^*$, ${p_y}^*$, and p_z represent the zero-centered PDFs in the measurement, median-relative, and standard-score spatial representations, respectively. Notably, the * superscript centers the distribution at zero by subtracting the associated median value as, ${p_x}^* = p_x - m_x$ for each spatial representation. More importantly, Equation 2.6 ensures that the hierarchal relationship between control function g_z, PDF p_z and CDF c_z in the standard-score space consistently projects into the measurement space or median-relative space. Hence, while the control function only mathematically exists in the standard-score space, the projection of the resulting PDF p_z simultaneously defines the probability of occurrence in all spatial representations.

Table 2.1 introduces the transformations for continuous zero-centered PDFs between each spatial representation. Transformations of the discrete data are accomplished using the median and standard deviation statistics and denoted in the "Magnitude" column of Table 2.1 where, x_i, y_i, and z_i represent the discrete data measurements in their respective spatial representations. The "PDF" and "derivative" columns introduce variable transformations that ensure conservation of probability within the CDFs in each spatial representation.

Table 2.1: Data transformations between each spatial orientation.

Space	Magnitude	PDF	Derivative
x	x_i	$p_x{}^* = \frac{1}{m_{x,i}} p_y{}^* = \frac{1}{\sigma_{x,i}} p_z$	$dx = m_{x,i} dy = \sigma_{x,i} dz$
y	$y_i = \frac{x_i}{m_{x,i}} = \frac{\sigma_{x,i}}{m_{x,i}} z_i + 1$	$p_y{}^* = \frac{m_{x,i}}{\sigma_{x,i}} p_z = m_{x,i} p_x{}^*$	$dy = \frac{\sigma_{x,i}}{m_{x,i}} dz = \frac{1}{m_{x,i}} dx$
z	$z_i = \frac{y_i - 1}{\frac{\sigma_{x,i}}{m_{x,i}}} = \frac{x_i - m_{x,i}}{\sigma_{x,i}}$	p_z	$dz = \frac{m_{x,i}}{\sigma_{x,i}} dy = \frac{1}{\sigma_{x,i}} dx$

Table 2.2 introduces the lower bound, central-tendency, and upper bound for parametric PDFs in each spatial representation. To reiterate, the probability of occurrence between the lower, central, and upper bounds within each spatial representation is retained, which implies conservation of probability. There is a 50-percent chance that data exist between the lower bound and central tendency, $\int_0^{m_x} p_x \, dx = \int_0^1 p_y \, dy = \int_{-\frac{m_{x,i}}{\sigma_{x,i}}}^0 p_z \, dz = \frac{1}{2}$; and there is 100-percent chance that data exist between the lower and upper bounds. Note that the lower bound of the standard-score space is dependent upon the median and standard deviation statistics and the central tendency of the measurement space is dependent upon the median. Therefore, evaluating the distribution solely in the standard-score space and projecting it into the measurement space could introduce measurement bias for processes where the median and standard deviation change with respect to time. In essence, the median-relative space evaluates the CDF using two constant reference points with a predefined probability, which alleviates this concern. This work identifies the median-relative space as an unbiased estimate of probability that projects into the measurement space and standard-score space. Using this spatial representation to evaluate all aspects of a PDF, including the mean statistic, allows the median and standard deviation to change without recursively influencing this interpretation of the PDF.

Table 2.2: Data boundaries in each spatial orientation.

	x_i	y_i	z_i
Lower Bound	$x_0 = 0$	$y_0 = 0$	$z_0 = -\frac{m_{x,i}}{\sigma_{x,i}}$
Central Tendency	m_x	$m_y = 1$	$m_z = 0$
Upper Bound	$x_{max} = \infty$	$y_{max} = \infty$	$z_{max} = \infty$

2.1.4. THE CONTROL FUNCTION

Equation 2.4 introduces the control function as the lognormal derivative of a continuous PDF. The nature of the control function defines the shape, or relative frequency, of the PDF in the standard-score space. There exist specific conditions where the control function will enforce the PDF to converge to a finite area on an unbounded interval, where that area can be scaled to unity. Specifically, the control function has to: 1) approach positive infinity as 'z' approaches negative infinity, $g_z \to +\infty$ as $z \to -\infty$; and, 2) approach negative infinity as 'z' approaches positive infinity, $g_z \to -\infty$ as $z \to +\infty$.

This work further qualifies the control function as a parametric representation that has the freedom and flexibility to match the shape of many discrete datasets. Parameterization of the control function such that it generally reproduces the shape of the histograms illustrated in Figure 2.1 is largely the topic of this chapter. Notably, the ability for the control function to accurately capture the information of the dataset is controlled by the number of parameters selected. Specifically, for the "Lenna" data increasing the number of parameters should provide more accurate reproduction of the measurement data; however, this data and its parametric PDF are only included as a proof of concept. Here, this section reiterates the progression of information necessary to estimate the nature of the control function. Specifically, the discrete data are expressed as a PMFs in the measurement space, with the median and standard deviation progressively transforming them into the median-relative space y and standard-score space z (see Table 2.1). The control function is only mathematically defined in the standard-score space, but it obviously embodies information related to the probability of occurrence on bounded intervals in the

measurement space x. In the next series of sections, this chapter introduces the parametric nature of the control function and progressively communicates the information necessary to parametrize it, starting with the ubiquitous normal distribution.

2.1.4.1. The Normal Distribution

Equation 2.7 introduces a first-order control function g_z which produces a normal distribution for a specific parameterization:

$$\begin{aligned} g_z &= -[\alpha_1 + \alpha_2 z] \\ p_z &= \exp\left(-\left[\alpha_0 + \alpha_1 z + \frac{\alpha_2}{2} z^2\right]\right) \end{aligned} \tag{2.7}$$

Where, α_0 is the constant of integration, and α_1 and α_2 represent the control function parameters. Note that the control function and PDF are essentially polynomial series with respect to the standard-score variable z. Moreover, the control function in Equation 2.7 has a negative slope given by $-\alpha_2$ with intercept $-\alpha_1$. Therefore, this relationship has the necessary properties of $g_z \to +\infty$ as $z \to -\infty$ and $g_z \to -\infty$ as $z \to +\infty$ to enforce convergence to unit area. Appendix A.2 shows that setting control function parameters to be $\alpha_1 = 0$ and $\alpha_2 = 1$ produces the normal distribution, which relegates the constant of integration to be $\alpha_0 = -\ln\left|\sqrt{1/2\pi}\right|$. Table 2.3 summarizes the control function parametrization for the normal distribution and provides the definition of p_z from Equation 2.7.

The control function parameter α_1 in Equation 2.7 serves to shift the PDF along the z-axis while maintaining unit area. Consequently, the general form of the linear control function $g_z = -\alpha_1 - \alpha_2 z$ could present useful properties for fitting the shape of the histograms shown in Figure 2.1. Hereafter, this root polynomial will serve as a basis for generating PDFs reminiscent of the normal distribution, but with shape attributes that more accurately reflect measurement data. In the next sections, this chapter extends this interpretation beyond the normal distribution to explore parametrization of the control function that has the freedom and flexibility to achieve this objective.

Table 2.3: Parametrization of the normal distribution and α_2 family of curves.

Historical Solution (Polar Coordinates)	**General Parametric Solution (See Appendix A.2)**
$\alpha_0 = -\ln\left\lvert\sqrt{\frac{1}{2\pi}}\right\rvert$ $\alpha_1 = 0$ $\alpha_2 = 1$	$\alpha_0 = -\ln\left\lvert\sqrt{\frac{\alpha_2}{2\pi}}\right\rvert$ $\alpha_1 = 0$ $0 < \alpha_2 < \infty$
$p_z = \exp\left(\ln\left\lvert\sqrt{\frac{1}{2\pi}}\right\rvert\right)\exp\left(-\frac{1}{2}z^2\right)$, the normal distribution	

2.1.4.2. α_2 Family of Curves

This section examines how the shape of the PDF in the standard-score space is governed by a linear control function with a vertical intercept of zero given by $\alpha_1 = 0$, but with $0 < \alpha_2 < \infty$. Using a change of variable, the tangent function transforms the α_2 parameter into an angular slope measured in degrees. Given that α_2 is the critical parameter allowing the control function to generate a PDF with unit area, this parameter can influence the resulting shape of the PDF resulting in the "α_2 family of curves." Equation 2.8 introduces the root polynomial control function for the α_2 family of curves:

$$g_z = -\left[\alpha_1 + \tan\left(\frac{\alpha_2\pi}{180}\right)z\right]$$
$$p_z = \exp\left(-\left[\alpha_0 + \alpha_1 z + \frac{1}{2}\tan\left(\frac{\alpha_2\pi}{180}\right)z^2\right]\right) \tag{2.8}$$

Equation 2.8 purposefully uses degrees instead of radians for two reasons. First, it is more intuitive. Second, converting the slope into a measure of degrees constrains α_2 to exist between $0° < \alpha_2 < 90°$, instead of using radians which is unbounded.

Figure 2.3 shows that the α_2 family of PDFs are bounded by familiar functions, with the normal distribution being an intermediate case. Figure 2.4 depicts the corresponding control

functions for these PDFs with the following angular slopes: $\alpha_2 = 0^\circ$ produces a uniform distribution; $\alpha_2 = 45^\circ$ produces the normal distribution; and, $\alpha_2 = 90^\circ$ produces a Dirac Delta function. By progressively increasing the angle α_2 from $0° \rightarrow 90°$, both the left and right-hand side tails of the PDF become less prominent and the distribution becomes more peaked. Note that α_2 contributes to the symmetry of the PDF while α_1 shifts it along the z axis. To reiterate, these numerical examples enforce $\alpha_1 = 0$ to ensure the distribution is centered about the standard-score origin. Finally, Table 2.3 provides a general form defining the constant of integration for the α_2 family of curves to be $\alpha_0 = -\ln\left|\sqrt{\alpha_2/2\pi}\right|$.

In general, the α_2 family of curves does not have enough freedom and flexibility to reproduce the histogram data shown in Figure 2.1, which exhibit attributes of being asymmetric, shifted, tail-weighted, and even multi-modal. Consequently, this approach extends the root polynomial control function in Equation 2.8 with additional polynomial or Fourier terms to adequately replicate the shape of these histograms.

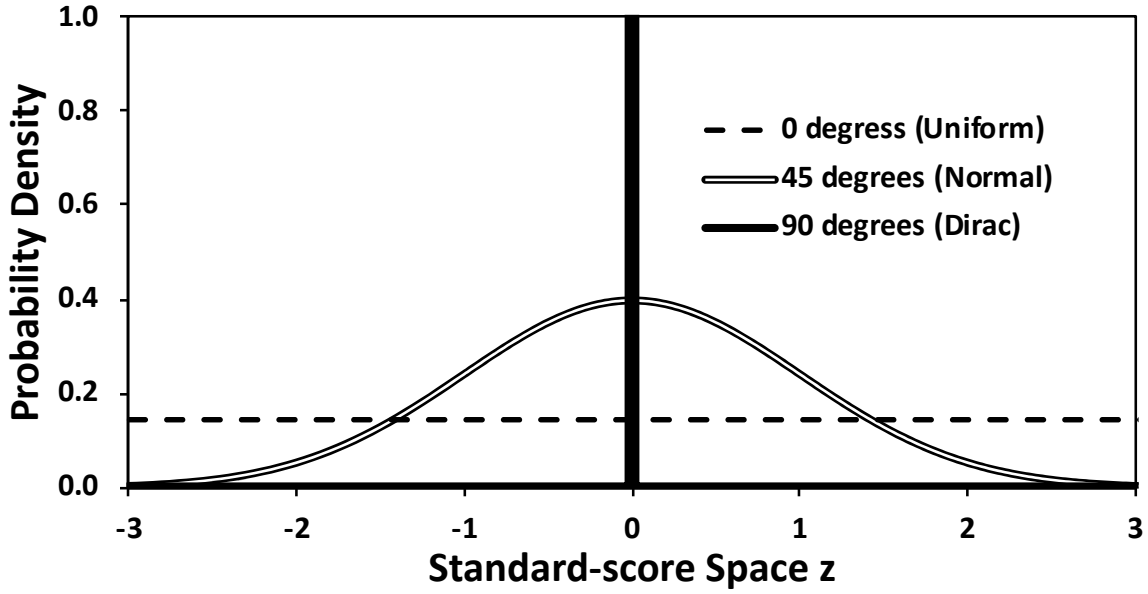

Figure 2.3: Bounding probability density functions for the α_2 family of curves.

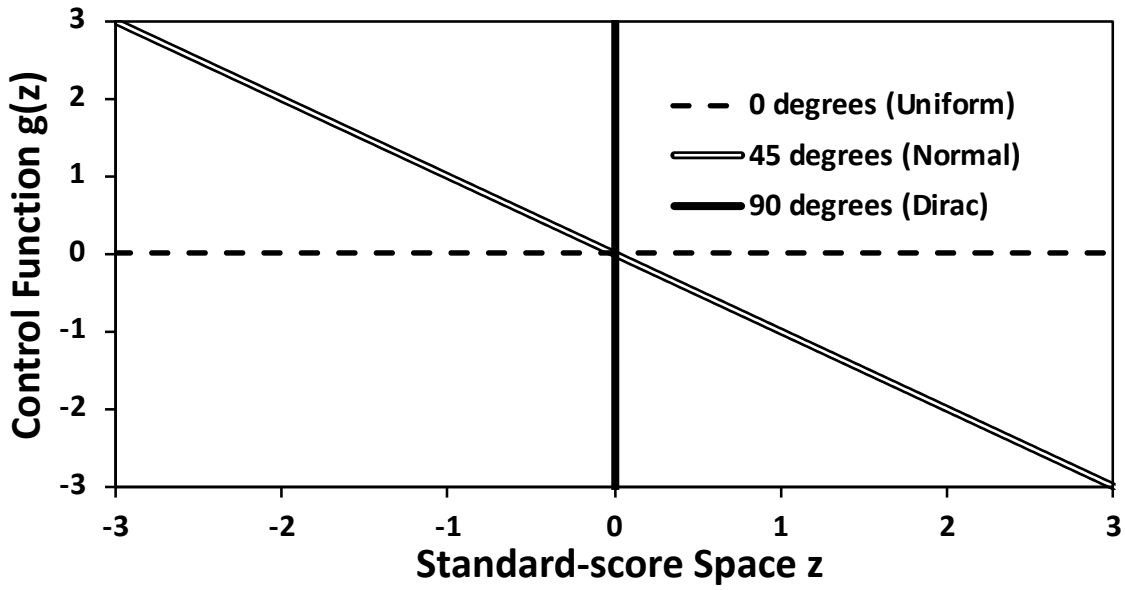

Figure 2.4: Corresponding control functions for bounding PDFs of the α_2 family of curves.

2.1.4.3. Polynomial Series Extension

Figure 2.1 shows the water consumption, hydraulic conductivity, and S&P 500 distributions that are unimodal, shifted, asymmetric, and tail-weighted. In order to replicate the shape of these histograms, this analysis extends the root polynomial control function to include additional terms in the series, as:

$$g_z = -\left[\alpha_1 + \tan\left(\frac{\alpha_2 \pi}{180}\right) z + \sum_{n_z=1}^{N_z} \alpha_{n_z+1} z^{n_z+1}\right] \tag{2.9}$$

Where, α_{n_z} is the parametric constant, n_z represents the order on the standard-score variable z, and N_z is the total order of the control function in the standard-score space. As before, the distribution is primarily defined by $0° < \alpha_2 < 90°$, which ultimately contributes to convergence. Experience suggests that terms subsequent to the root polynomial control function diminish in significance. Practical application of Equation 2.9 is limited to distributions that are unimodal but may be asymmetric, shifted, and tail-weighted. In general, odd polynomials $\alpha_{1,3,5,7\ldots}$ contribute to the asymmetry of the PDF, whereas even polynomials $\alpha_{2,4,6,8\ldots}$ contribute to the peakedness of the distribution. Finally, the integration constant α_0 must be defined to ensure the PDF has unit area.

Analytical integration techniques to evaluate closed-form expressions of α_0 for parametric PDFs may be intractable; therefore, numerical integration provides an alternative approach.

2.1.4.4. Modified-Fourier Series Extension

Figure 2.1 shows the Lenna light intensity histogram as a multi-modal distribution which clearly cannot be replicated by a simple polynomial series. To accommodate the wave-like nature of multiple peaks, this analysis extends the root polynomial of the α_2 family of curves to include a modified-Fourier series, $\mathcal{F}_{n_z,N_\mathcal{F}}$, as:

$$\begin{aligned} \mathcal{F}_{n_z,N_\mathcal{F}} &= \sum_{n_\mathcal{F}=0}^{N_\mathcal{F}} \upsilon_{n_z,n_\mathcal{F}} \sin\left(\psi_{n_z,n_\mathcal{F}} z + \varrho_{n_z,n_\mathcal{F}}\right), \qquad \mathcal{F}_{n_z,0} = 0 \\ g_z &= -\left[\alpha_1 + \mathcal{F}_{0,N_\mathcal{F}} + \tan\left(\frac{\alpha_2 \pi}{180}\right)\left\{1 + \mathcal{F}_{1,N_\mathcal{F}}\right\} z + \sum_{n_z=2}^{N_z} \mathcal{F}_{n_z,N_\mathcal{F}} z^{n_z}\right] \end{aligned} \tag{2.10}$$

Where, $N_\mathcal{F}$ is the total number of modified-Fourier sinusoidal waves and n_z represents the order of the standard-score variable z. In Equation 2.10, three constants, $\upsilon_{n_z,n_\mathcal{F}}$, $\psi_{n_z,n_\mathcal{F}}$ and $\varrho_{n_z,n_\mathcal{F}}$, parameterize each modified-Fourier series wave. Similar to the polynomial series extension, the control function is primarily controlled by $0° < \alpha_2 < 90°$. However, this approach allows for a period function to supplement the angular slope α_2 along the horizontal axis. This permits the modified-Fourier series greater freedom for fitting oddly-shaped and even multi-modal datasets, as demonstrated in the application section. The implications of the polynomial and modified-Fourier series extensions are described in the application section.

2.1.5. MEDIAN-RELATIVE SPACE

Previously, Table 2.1 introduced the dimensionless median-relative space that expresses the probability of occurrence for the discrete data and parametric PDFs without bias from the median and standard deviation statistics. This permits comparison of seemingly disparate datasets by transforming the shape of the distribution from the standard-score space to the median-relative space: $p_y = \frac{m_{x,i}}{\sigma_{x,i}} p_z$. Minimizing the mean-squared error between the CDF c_y and CMF c_{y,k_1} in the median-relative space provides this analysis with a robust objective function for parameterization of the control function. Clearly, the statistics, hierarchical integrations that promote conservation

of probability, and functional relationships between the x, y and z spatial orientations are prominent features of this work.

2.1.5.1. Data Culling

All data measurements naturally arise in the measurement space characterized as x_i. When transformed into median-relative y_i or standard-score z_i form, the natural upper bound remains an infinitely large measurement. However, very large magnitude measurements may be symptomatic of either excessive measurement error or perhaps observations from another distinct population. Population outliers can potentially bias this evaluation of the median and standard deviation, as well as the parameters within the control function given their reliance on the standard-score space. The focus of this section is to predefine a consistent upper bound in the median-relative space y_{max} that removes potential population outliers from discrete datasets and analogously applies to any dataset regardless of location or scale.

This analysis only consider datasets that have discrete data within the range $x_0 < x < x_{max}$, and hence are comprised of real, non-zero, and positive measurements (See Figure 2.1). This range reflects values on the median-relative axis on the interval $y_0 < y < y_{max}$. The first position is the measurement-space origin denoted by a zero-magnitude measurement $x_0 = y_0 = 0$, which simply transforms into the origin of the median-relative space. Before parameterizing a PDF to reflect the shape of the histogram, this approach discards all data greater than the culling threshold $y_i > y_{max}$. Herein, this approach applies a heuristic by selecting y_{max} to be a multiple of "$m_{y,i} = 1$" and then apply the same value to each histogram. A consistent data culling threshold y_{max} ensures data are retained to the same degree for the disparate datasets, regardless of their scale in the measurement space.

Data culling can potentially introduce recursive adjustments in the median and culling threshold and hence mapping of $x_i \leftrightarrow y_i \leftrightarrow z_i$. However, the median is seemingly insensitive to the low frequency at which extremely large erroneous measurements occur, and hence datasets may require significant culling before observing changes to the median. In contrast, the standard deviation is quite sensitive to high magnitude outliers. Therefore, data culling is a necessary step to generate the correct estimation of $\sigma_{x,i}$ providing accurate and consistent mapping between the continuous representation of the discrete data between each spatial transformation $x \leftrightarrow y \leftrightarrow z$ as shown on Table 2.1.

2.1.5.2. Objective Function

Upon culling population outliers, the framework minimizes an objective function using a MSE approach to estimate parameters within the polynomial and/or modified-Fourier series control functions. The objective function penalizes the difference between the CDF and CMF as:

$$MSE_{c,y} = \frac{1}{N_k}\sum_{k=1}^{N_k}\left[c_y - c_{y,k}\right]^2 \tag{2.11}$$

Where, N_k represents the number of bins in the analysis. Minimizing the MSE in Equation 2.11 creates a parametric PDF that reproduces the shape of the histogram data. This process relies on the hierarchal relationship between the control function, PDF, and CDF to ensure the parametric PDF correctly reproduces the PMF for all reasonable measurements along each spatial representation, concurrently.

A key insight is that the median-relative space allows this analysis to produce equally-spaced probability interval bins k, while supporting a general algorithm that allows application of the objective function to many datasets as disparate as those on Figure 2.1. Additionally, these bins are defined independent of, and prior to, the control function parametrization.

2.1.5.3. The Mean Statistic

This section demonstrates that the mean of the distribution is fully defined by a combination of median, standard deviation, and control function statistics. The probability-weighted mean for a PDF p_z in the standard-score space is defined on some interval as:

$$\mu_z = \int_{z_0}^{z_{max}} z p_z \, dz \tag{2.12}$$

Where, μ_z represents the mean statistic in the standard-score space z, and $z_0 = -\frac{m_{x,i}}{\sigma_{x,i}}$ while $z_{max} = \frac{y_{max}-1}{\frac{\sigma_{x,i}}{m_{x,i}}}$ where $z_0 < z < z_{max}$, The mean statistic occupies a single position on the distribution and can be mapped through each spatial orientation. Equation 2.13 demonstrates this mapping of the mean statistic for the parametric PDF between $x \leftrightarrow y \leftrightarrow z$. Furthermore, the mean

is entirely defined by the median, standard deviation, and control function as follows and is derived in Appendix A.3.

$$\mu_x = m_{x,i}\mu_y = m_{x,i} + \sigma_{x,i}\mu_z \quad \Longrightarrow \quad \mu_x = m_{x,i} + \sigma_{x,i}\int_{z_0}^{z_{max}} \left[z \exp\left(\int g_z \, dz \right) \right] dz \tag{2.13}$$

The arithmetic mean of the discrete dataset can be compared to the probability-weighted mean of the corresponding parametric PDF to empirically evaluate its goodness of fit. The mean statistic on its own is not sufficient to characterize goodness of fit because there are an infinite number of distributions that could result in the same mean statistic but with varying shapes. Therefore, the mean statistic is not included in the objective function; however, the mean statistic of the parametric PDF will naturally gravitate toward the arithmetic mean of the discrete dataset as a consequence of minimizing the objective function in Equation 2.11.

To proceed, this analysis considers a comparison of the mean statistic by first empirically defining the arithmetic mean of the dataset in the measurement space. Equation 2.14 relates the arithmetic mean in the measurement space $\mu_{x,i}$ to an analogous value in the median-relative space $\mu_{y,i}$.

$$\begin{gathered} \mu_{y,i} = \frac{1}{N_i}\sum_{i=1}^{N_i} y_i = \frac{1}{m_{x,i}N_i}\sum_{i=1}^{N_i} x_i \\ \because \mu_{x,i} = \frac{1}{N_i}\sum_{i=1}^{N_i} x_i\,, \qquad \therefore \mu_{y,i} = \frac{1}{m_{x,i}}\mu_{x,i} \end{gathered} \tag{2.14}$$

A novel contribution of this analysis defines the median-relative space arithmetic mean $\mu_{y,i}$ to be a transformation of the measurement space arithmetic mean: $\mu_{y,i} = \frac{1}{m_x}\mu_{x,i}$. This ratio is unity for a normal distribution and increases in value as the distribution becomes progressively tail-heavy.

Using a MSE approach, Equation 2.15 provides an independent measure to verify the parametrization of the control function fitting the CDF c_x to the CMF c_{x,k_1}.

$$MSE_{\mu,y} = \left[\mu_{y,i} - \mu_y\right]^2 \tag{2.15}$$

Equation 2.15 is analogous to the objective function, but instead constitutes how effectively the control function selection expresses the continuum behaviour of the collective data. Minimizing the objective function given by Equation 2.11 constrains the continuous PDF to be nearly identical to the PMF, given an appropriate control function. Notably, the objective function in Equation 2.11 does not guarantee nor even suggest that Equation 2.15 will represent a global minimum. The applications considered herein show that selecting an appropriate control function results in commensurate accuracy for the $MSE_{c,y}$ and $MSE_{\mu,y}$.

At this point, it is apparent that the median, standard deviation, and control function parametrization embody all of the information necessary to reproduce the discrete dataset as a PDF. In other words, they compress all information pertaining to the distribution into a reduced set of scalar values. The median-relative space guarantees a constant frame of reference for evaluating the scale and shape of a PDF and provides the foundation for viewing the mean statistic as a solution to an advection-dispersion problem (see Equation 2.13).

2.1.6. DEGREES OF FREEDOM ANALYSIS

Here, a degrees of freedom analysis is used to evaluate the effectiveness of compressing the histogram data into a PDF using the median, standard deviation, and control function parametrization. Assuming these statistics represent one degree of freedom each, the parametric compression of many datasets can be evaluated in the median-relative space using the relationships in Table 2.4. A zero-centered distribution is an example where evaluating degrees of freedom within the median-relative space is not possible, because this requires division by zero.

Table 2.4: Degrees of freedom analysis.

Description	Measure	Degrees of Freedom
Arithmetic Mean	$\mu_{y,i} = \frac{1}{N_i}\sum_{i=1}^{N_i} y_i$	$N_\mu = 1$
Probability-Weighted Mean	$\mu_y = \int_{y_0}^{y_{max}} y p_y \, dy$	
Median	$m_{y,i} = 1$	$N_m = 1$
Standard Deviation	$\sigma_{y,i} = \frac{\sigma_{x,i}}{m_{x,i}}$	$N_\sigma = 1$
Control Function	$g_z(\alpha, \upsilon, \psi, \varrho \ldots N_{CF})$	N_{CF}
PDF	$p_y = \frac{m_{x,i}}{\sigma_{x,i}} \exp\left(\int g_z \, dz\right)$	$N_{PDF} = N_\mu + N_m + N_\sigma + N_{CF}$
Parametric Compression	$\alpha, \upsilon, \psi, \varrho$	$N_{PC} = N_i - N_{PDF}$
Discrete Data	x_i	N_i
Compression Efficiency	$\mathcal{C}$	$\frac{N_i - N_{PDF}}{N_i} \times 100\%$

Note: N_i represents the data remaining in the analysis after culling occurs.

2.2. APPLICATION

This section applies the methodology pertaining to control function theory to the four histograms shown on Figure 2.1, which represent datasets from economics, engineering, finance, and image analysis. The diversity of data sources is meant to strengthen this empirical demonstration and exhibit the generality of this approach. To begin, the heuristic approach can be summarized as:

1. Evaluate the median $m_{x,i}$ and standard deviation $\sigma_{x,i}$ of the discrete dataset x_i.
2. Transform the discrete data x_i into the median relative space using $y_i = \frac{x_i}{m_{x,i}}$.
3. Perform data culling on y_i for all datasets with the predefined threshold y_{max}. This threshold may be adjusted to balance the need to both minimize the amount of culled data, and also minimize the distortion of large magnitude outliers on $m_{x,i}$ and $\sigma_{x,i}$.
4. Finalize the standard deviation $\sigma_{x,i}$ and arithmetic mean $\mu_{x,i}$ of the culled data.
5. Create discrete bins k within the culled dataset x_i to generate histograms $h_{x_{k-1}<x_i<x_k}$and the probability of occurrence $p_{x,k}$ within each bin.
6. Transform the PMF $p_{x,k}$ into a CMF $c_{x,k_1}$and then map it into the median-relative space, $c_{x,k_1} \rightarrow c_{y,k_1}$.
7. Choose an appropriate control function to generate a CDF c_z and then map $c_z \rightarrow c_y$.
8. Incrementally add terms to the control function extension and use an optimization strategy that adjusts parameters within the control function g_z to minimize the objective function such that $c_{y,k_1} \cong c_y$.
9. Evaluate the appropriateness of the control function by comparing the arithmetic mean $\mu_{y,i}$ to the mean statistic μ_y in the median-relative space.

Step 3 estimates the median and standard deviation, while culling data from the water consumption and S&P 500 datasets using $0 < y_i < 4$ as the range for inclusion, with all details of this data culling exercise summarized on Table 2.5. Both the hydraulic conductivity and Lenna datasets require no culling as all data exist on the interval $0 < y_i < 4$. Note that y_i is dimensionless and hence no units are reported for the various datasets. The water consumption data has 162 data points beyond the culling threshold $y_{max} = 4$ that have a disproportionate influence on the standard deviation of the distribution. Including these data points increases the standard deviation

from $2.57\text{x}10^1$ to $2.93\text{x}10^1$, which is an increase of approximately 15% for data reflecting less than 1% of the population. Failure to cull this data would bias the parameter estimation of the control function when enforcing $c_{y,k_1} \cong c_y$. The S&P 500 data has 8 points beyond the threshold $y_{max} = 4$ that have a disproportionate influence on the arithmetic mean of the distribution. Including these data points increases the arithmetic mean from $3.40\text{x}10^1$ to $3.77\text{x}10^1$, which is an increase of approximately 11% for data reflecting less than 2% of the population. Variation in the mean statistic suggests the culled data has undue influence on the shape of the distribution, because the median and standard deviation remain relatively constant.

Table 2.5: Summary of statistics for the four disparate datasets.

	Water Consumption	**Hydraulic Conductivity**	**S&P 500 08/21/2009 Index**	**Lenna Light Intensity**
Data and Statistics				
Total Measurements	22,509	720	499	262,144
Data Points Culled	162	0	8	0
Analysis Data Points (N_i)	22,347	720	491	262,144
Median $(m_{x,i})$	$4.00\text{ x}10^1$	$9.93\text{ x}10^{-3}$	$3.07\text{ x}10^1$	$1.29\text{ x}10^2$
Standard deviation $(\sigma_{x,i})$	$2.57\text{ x}10^1$	$5.64\text{ x}10^{-3}$	$1.97\text{ x}10^1$	$4.81\text{ x}10^1$
Arithmetic Mean $(\mu_{x,i})$	$4.45\text{ x}10^1$	$1.11\text{ x}10^{-2}$	$3.40\text{ x}10^1$	$1.23\text{ x}10^2$
Polynomial Series Extension				
PDF (N_{PDF})	8	8	8	
Parametric Compression (N_{PC})	22,339	712	483	n/a
Compression Efficiency $(\mathcal{C})$	99.28%	98.89%	98.37%	
Modified-Fourier Series Extension				
PDF (N_{PDF})	8	8	8	17
Parametric Compression (N_{PC})	22,339	712	483	262,127
Compression Efficiency $(\mathcal{C})$	99.28%	98.89%	98.37%	99.99%

The culled discrete data representing the water consumption, hydraulic conductivity, and S&P 500 index sources were arranged into 16 discrete bins of size $\Delta y = 0.25$ within the median-relative space over the interval $0 \leq y \leq 4$. Given the multi-modal nature of the Lenna histogram,

the analysis implements 86 discrete bins of size $\Delta y \cong 0.0234$ over the interval $0 \leq y \leq 4$ to resolve the PMF as a PDF. After culling the population outliers, the analysis calculates the probability of occurrence within the aforementioned Δy intervals. Next, these probabilities were summed into a CMF c_{x,k_1}, and then mapped to c_{y,k_1} using the median statistic $m_{x,i}$. The probability of occurrence for intervals within each application are itemized in Appendix A.4.

Selecting a control function $g_z(\alpha, \upsilon, \psi, \varrho \ldots N_{CF})$ from Equations 2.8 and 2.9 allow this analysis to replicate the CMF c_{y,k_1} of each dataset as a CDF c_y. This step requires parameter estimation of $\alpha, \upsilon, \psi, \varrho$ within the control function for either the polynomial or modified-Fourier series extensions. This application considers both the polynomial and modified-Fourier series extensions for the unimodal datasets and applies the modified-Fourier series extension for both the unimodal and multi-modal datasets. The parameter α_0 in Equations 2.6 and 2.7 are similarly present in the standard-score PDFs for each application and numerical integration constrains α_0 to ensure unit area beneath each PDF. This scaling process ensures conservation of probability for each application. To this end, Simpson's Rule is applied within the standard-score space using a discretization of $\Delta z = 0.02$ on the interval $-\frac{m_{x,i}}{\sigma_{x,i}} < z < \frac{y_{max}-1}{\frac{\sigma_{x,i}}{m_{x,i}}}$, while concurrently changing the control function parameters to minimize the objective function in Equation 2.10 for each application. Table 2.6 introduces characteristic control functions that parametrically reproduce each dataset.

As mentioned earlier, the analysis applies the exponential polynomial and modified-Fourier series parameterization for each of the water consumption, hydraulic conductivity, and S&P 500 index datasets. Notably, these parameterizations require the same number of terms to accurately reproduce the datasets through the polynomial and modified-Fourier series extensions of the control function. The polynomial series extension for these three datasets applies two additional terms beyond the root polynomial control function. Additionally, the modified-Fourier series extension for these three datasets applies one sinusoidal wave with three additional parameters beyond the root polynomial control function. The parametric compression N_{PC} for both the polynomial and modified-Fourier series extensions are identical and are listed on Table 2.5. Note that $\varrho_{0,1}$ for the water consumption, hydraulic conductivity, and S&P 500 index data is necessarily "zero" because the slope of the control function does not change from negative to

positive, thus there is no change in concavity. Hence, only $\upsilon_{0,1}$ and $\psi_{0,1}$ contribute to replicating discrete data as unimodal PDFs. This ensures that the compression efficiency $\mathcal{C}$ is identical for both control function extensions as applied to these unimodal distributions (see Table 2.5). Later this chapter discusses how the modal nature of the distribution is inherently linked to changes in concavity.

Figures 2.5 and 2.6 present results from the parameter estimation exercise for all four datasets. These figures depict the shape of the control function g_z and resulting PDF p_z in the standard score space. On Figure 2.5, note the stark difference between control functions that characterize unimodal and multi-modal distributions. Unimodal distributions express control functions that have a varying but negative slope across all z, but do not experience changes in concavity. In contrast, the modified-Fourier series control function for the Lenna dataset observes multiple changes in concavity, which roughly correspond to the peaks observed on the histogram in Figures 2.1.d and parametric PDF in Figures 2.2.d and 2.6. Qualitatively, this suggests that there is an innate link between control function concavity and the modal characteristics of the associated PDF. Notably, the functions g_z and p_z for the water consumption, S&P 500 and hydraulic conductivity datasets are visually indistinguishable between the polynomial and Fourier series approaches given the low $MSE_{c,y}$ obtained when minimizing Equation 2.10 for both approaches. Control function parameters that result from minimizing Equation 2.10 for each application are itemized on Tables 2.7 and 2.8.

Table 2.6: Exponential polynomial and modified-Fourier series control functions for each application.

Water Consumption, Hydraulic Conductivity, and S&P 500 Index Data	
Polynomial Series	$g_z = -\left[a_1 + \tan\left(\frac{\alpha_2\pi}{180}\right)z + \alpha_3 z^2 + \alpha_4 z^3\right]$
Modified-Fourier Series	$\mathcal{F}_{0,1} = \upsilon_{0,1}\sin\left(\psi_{0,1}z + \varrho_{0,1}\right)$ $g_z = -\left[a_1 + \mathcal{F}_{0,1} + \tan\left(\frac{\alpha_2\pi}{180}\right)z\right]$
Lenna Intensity Parametric Control Function	
Modified-Fourier Series	$\mathcal{F}_{0,2} = \upsilon_{0,1}\sin\left(\psi_{0,1}z + \varrho_{0,1}\right) + \upsilon_{0,2}\sin\left(\psi_{0,2}z + \varrho_{0,2}\right)$ $\mathcal{F}_{1,2} = \upsilon_{1,1}\sin\left(\psi_{1,1}z + \varrho_{1,1}\right) + \upsilon_{1,2}\sin\left(\psi_{1,2}z + \varrho_{1,2}\right)$ $g_z = -\left[\alpha_1 + \mathcal{F}_{0,2} + \tan\left(\frac{\alpha_2\pi}{180}\right)\{1 + \mathcal{F}_{1,2}\}z\right]$

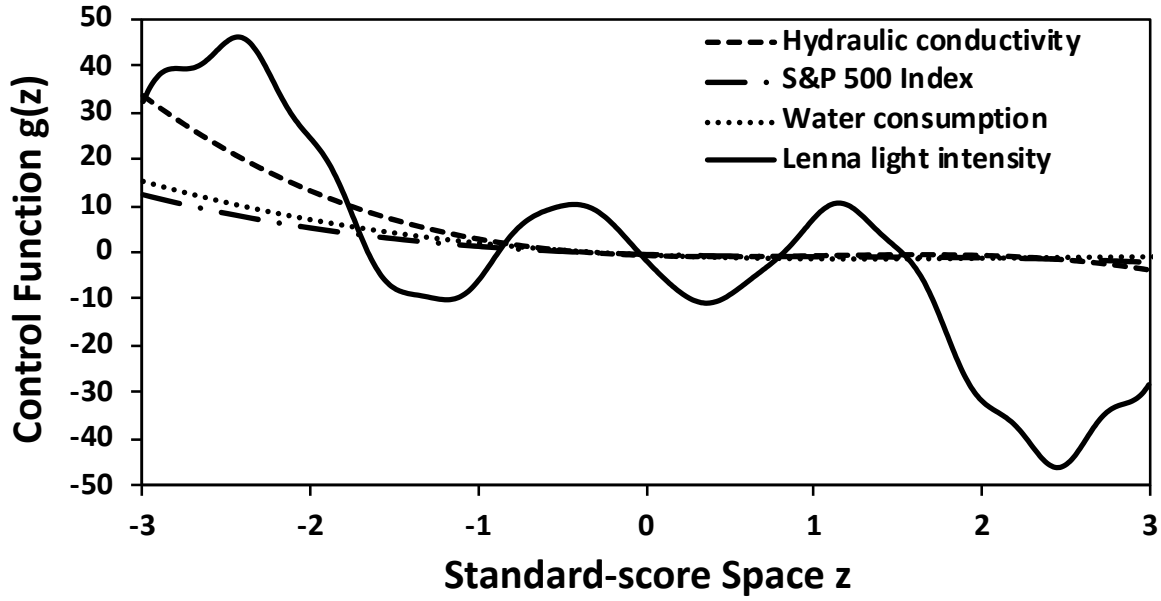

Figure 2.5: Control function visualization for each histogram.

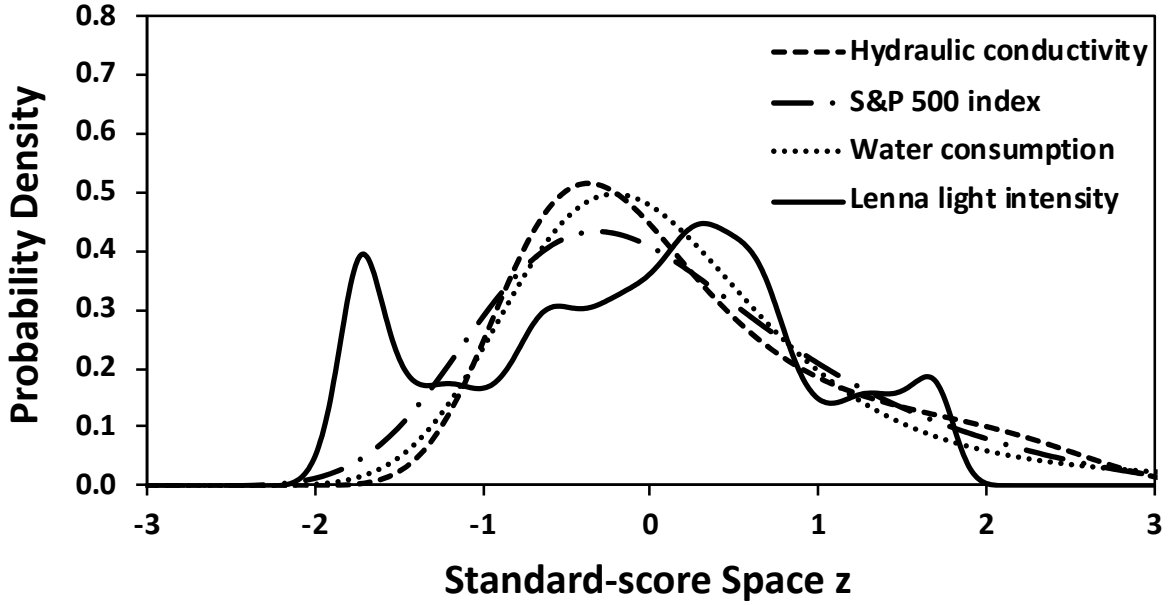

Figure 2.6: Standard-score PDF visualization for each histogram.

Table 2.7: Parameterization for the water consumption, hydraulic conductivity, and S&P 500 datasets.

	Water Consumption	Hydraulic Conductivity	S&P 500 Index 08/21/2009
CDF Bins	16	16	16
Polynomial Series Extension			
$MSE_{\mu,y}$	6.77 x10^{-5}	2.24 x10^{-5}	1.02 x10^{-4}
$MSE_{c,y}$	8.08 x10^{-6}	2.61 x10^{-5}	1.04 x10^{-5}
α_0	-0.7344	-0.8061	-0.8950
α_1	0.3625	0.7014	0.3612
α_2	57.2188°	50.7525°	43.3229°
α_3	-0.8475	-1.7144	-0.6196
α_4	0.1249	0.5502	0.1653
Modified-Fourier Series Extension			
$MSE_{\mu,y}$	4.78 x10^{-6}	3.40 x10^{-5}	1.83 x10^{-4}
$MSE_{c,y}$	9.23 x10^{-5}	3.12 x10^{-4}	7.71 x10^{-5}
α_0	1.2730	1.1166	1.1111
α_1	0.0584	0.0421	0.1048
α_2	26.3296°	31.6049°	29.1762°
$\upsilon_{0,1}$	0.9582	1.0539	0.4359
$\psi_{0,1}$	1.6781	2.3394	1.7898
$\varrho_{0,1}$	0.0000	0.0000	0.0000

Table 2.8: Parameterization for the Lenna light intensity dataset.

Control Function Parameter	Lenna Light Intensity
CDF Bins	86
$MSE_{\mu,y}$	1.77 x10^{-5}
$MSE_{c,y}$	8.91 x10^{-6}
α_0	-2.1826
α_1	0.4213
α_2	82.0695
$\upsilon_{0,1}$	5.2015
$\psi_{0,1}$	4.0151
$\varrho_{0,1}$	0.2116
$\upsilon_{0,2}$	-0.1827
$\psi_{0,2}$	1.0838
$\varrho_{0,2}$	-2.8071
$\upsilon_{1,1}$	-1.6093
$\psi_{1,1}$	2.5347
$\varrho_{1,1}$	-1.4543
$\upsilon_{1,2}$	-0.1561
$\psi_{1,2}$	11.6571
$\varrho_{1,2}$	-5.1059

Figure 2.7 provides a direct comparison between the discrete and continuous CDFs after achieving the minimum objective function in Equation 2.10. Low $MSE_{c,y}$ values for each application suggest that the control function accurately replicates the shape of each dataset over the entire range of $0 \leq y \leq 4$. Figure 2.7 presents a comparison of the CDF and CMF in the measurement space. These parametric CDFs correspond to the parametric PDFs presented in Figure 2.2 for each application.

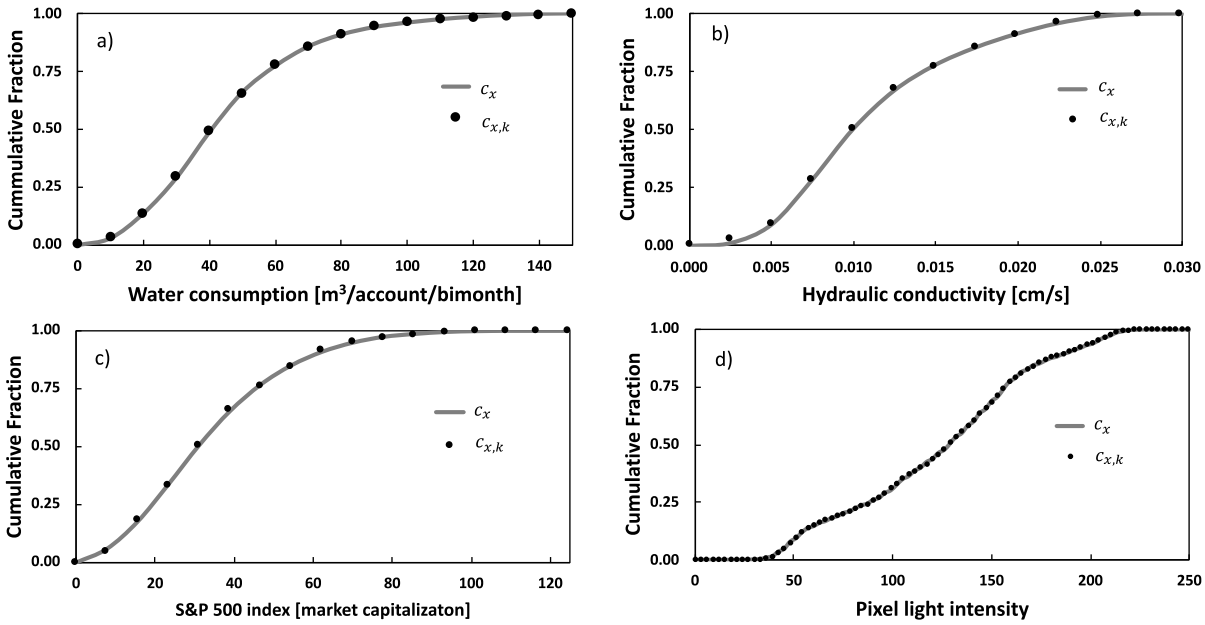

Figure 2.7: Visualization of the CMF and CDF for the optimally parameterized PDFs. This figure includes results from minimizing the objective function such that $c_{y,k_1} \cong c_y$ and $c_{x,k_1} \cong c_x$.

Four issues merit discussion regarding the resulting fit between the discrete datasets and their continuous parametric counterparts. First, when conducting numerical integration to estimate α_0, numerical error may impact control function parameterization, and hence the ability to fit the shape of discrete data while minimizing $MSE_{c,y}$. Second, additional control function terms could have more accurately reproduced each dataset. In particular, this would significantly improve parameterization of the Lenna light intensity CDF in order to replicate each individual peak shown on Figures 2.2.d and 2.6. The ability to preselect the number of Fourier terms in advance of minimizing the objective function requires additional conceptualization beyond what is presented in this framework. Third, the analysis estimates the median and standard deviation statistics directly from the discrete data, hence these statistics are subject to measurement error. Additionally, the parametric fit of the control may introduce further error from reproducing the discrete data as a smooth function. Fourth, while all CDFs have $c_y = 0$ when $y = 0$, the water consumption and S&P 500 CDFs rapidly rise to a non-zero value of c_y upon the first bin k. This occurs because there is non-zero probability of recording the least non-zero measurement magnitude for these data sets.

A key outcome of this empirical assessment is that the arithmetic mean of each dataset $\mu_{x,i}$ is similar to the probability-weighted mean of the parametric PDF μ_x as established by small values of $MSE_{\mu,y}$ (see Tables 2.7 and 2.8). To reiterate, minimizing Equation 2.10 indirectly enforces the arithmetic mean of the discrete dataset to be nearly equal to the probability-weighted mean of the PDF for an appropriate control function. The S&P 500 application produces the largest observed error, $MSE_{\mu,y} = 1.83 \text{ x} 10^{-4}$ while applying the modified-Fourier series control function. Given that data in the median-relative space is dimensionless, this error is only $\sqrt{1.83 \times 10^{-4}}$ as a percentage of the median $m_{x,i}$. Additionally, and in reference to the exponential polynomial application of the water consumption distribution where $m_{x,i} = 40\ m^3/account/period$, the error of 6.77×10^{-5} amounts to approximately 0.80% of the median water consumption. This translates into a total error of $7{,}027\ m^3/account/period$ during the July/August 2007 billing period for all 22,347-single-family residential accounts. In summary, this chapter concludes via this empirical application that the control function is able to delineate many PDFs using the median and standard deviation statistics as defined by each dataset. Furthermore, the mean statistic is entirely defined by the median, standard deviation, and control function. This work clearly demonstrates the reproducibility of this method through its ability to evaluate a wide variety of systems relevant to fields as disparate as economics, engineering, finance, and image analysis.

Given that the four applications comprise hundreds to thousands of data points, the fact that the continuum-level information can be replicated with a few parameters in a continuously differentiable control function and PDF implies "data compression". Table 2.5 itemizes the compression efficiency $\mathcal{C}$ for each set of parameters, with a minimum value of 98.37% for the S&P 500 dataset.

Future application of this technique will serve to use concepts of probability, time-dependence and spatial reference information to forecast how the evolution of the median, standard deviation, and control function predictably influence the mean statistic. This latter point is predicated on the idea that a consistent set of control function parameters can relate causality between influential processes and the shape of the distribution. The notion of combining probability, time-dependence, and ambient conditions provides the foundation for viewing both the PDF and its mean statistic, as expressed in Equation 2.12, as a solution to an advection-dispersion problem. This realization allows provides motivation to reimagine the advective-

dispersive process in the context of the statistical transformations in Table 2.1. Appendix A.5 carefully describes the motivation for viewing PDFs as the solution to an advective-dispersive process that ultimately results in Equation 2.16. The median represents the central tendency or bulk location of the distribution, while the standard deviation and standard-score PDF combine to characterize the scale and shape of the distribution. Therefore, this analysis contends that changes to the median, standard deviation, and standard-score PDF through the control function are commensurate to advection and dispersion. This provides the motivation for investigating systems that observe a continuous shift in the distribution of empirical results through probabilistic advection and dispersion using the following relationship:

$$\underbrace{p_x}_{\substack{continuum\\distribution}} = \underbrace{m_x}_{\substack{probabilistic\\advective\\process}} + \underbrace{\frac{1}{\sigma_x} \times p_z}_{\substack{probabilistic\\dispersive\\process}} \tag{2.16}$$

This interpretation of advection and dispersion may provide a path to deconstruct complex probabilistic processes into the simple concepts of location, scale, and shape. The ultimate goal of this interpretation would be to individually model these statistics and recombine them to reproduce, model, and potentially forecast how these complex processes will evolve through time.

2.3. CONCLUSIONS

This work demonstrates how a combination of statistics and calculus can characterize the relationships between the CDF, PDF, and control function. A motivation for this work is to compress discrete datasets without assuming a predefined shape for the distribution of measurements as an alternative to kernel density estimation. This framework evaluates the control function and combines it with median and standard deviation statistics to produce a continuously differentiable function that can replicate discrete datasets. Collectively, individual measurements of a system are compressed and expressed through the median, standard deviation, and control function, while the mean statistic represents the ensemble system behaviour. Specifically, this chapter builds theoretical and empirical evidence to demonstrate the ubiquitous nature of the median, standard deviation, and control function, which reflect measures of location, scale, and shape that uniquely define a distribution. Understanding how these conditions change through time

relative to ambient conditions, is paramount when forecasting how the mean statistic of a system will evolve. These revelations warrant future discussion to clearly identify the philosophical implications arising from this work. The following conclusions are drawn from this study:

1. Control function theory represents a form of kernel density estimation that relies upon hierarchical relationships between control function, PDF, and CDF to produce a smooth representation of otherwise discontinuous data, thus capturing all the information of a discrete dataset in a compressed functional form.
2. Histogram data representing water consumption, hydraulic conductivity, S&P 500, and photo light intensity are transformed between the measurement, median-relative, and standard-score spaces using the median and standard deviation statistics. The median-relative space divides the discrete data by its median value to produce a dimensionless dataset.
3. Collectively, individual measurements define the continuum condition of any system. The median represents the continuum location of the dataset, whereas the standard deviation represents the continuum scale of the dataset. The parametric control function provides an unambiguous link between the PMF and its continuous representation as a PDF. Therefore, the control function is continuum condition that represents the shape of the distribution.
4. A parametric control function results in a continuously differentiable PDF that can reproduce familiar distributions. The normal distribution is generated by a linear control function with a negative slope. Supplementary to the root polynomial control function, additional polynomial or Fourier terms can reproduce the attributes of asymmetric, tail-weighted, or multi-modal distributions.
5. Control function theory facilitates highly effective compression of discrete datasets by evaluating the continuum statistics quantifying the location, scale, and shape of a distribution. Application to datasets from engineering, finance, economics, and image analysis show a compression efficiency in excess of 98%.

This work provides a foundation for developing time-dependent, probabilistic relationships that characterize how the continuum statistics of a PDF and its resulting mean statistic will evolve through time and as a function of ambient conditions. The median-relative space guarantees a constant frame of reference for evaluating the scale and shape of the distribution, which in turn provides the foundation for viewing the PDF and its mean statistic as a solution to an advection-dispersion problem.

3. Advective-Dispersive Transport of a Probability Density Function: Model Development and Water Consumption Application

Water utilities need to accurately forecast residential water consumption so that they can adjust the unit price of water and balance revenues with expenses, enabling them to operate their system on a full-cost recovery basis. United States Environmental Protection Agency (EPA) documents a shift in policy goals over the past 20 years toward the financially sustainable operation and management of water utilities (EPA, 2003; 2005; 2006). In Ontario, Canada, the Water Opportunities and Conservation Act mandates that municipal water utilities develop sustainability plans for their water distribution services (MOE 2011). Hughes and Leurig (2013) suggest that changing water consumption habits has resulted in considerable revenue uncertainty, potentially sabotaging utility efforts to develop financially sustainable management practices. House-Peters and Chang (2011) and Donkor et al. (2011) provide a comprehensive review of the advances in methodologies for urban water forecasting and analysis that quantify trends in water consumption, such as: econometrics, agent-based, system dynamics, and artificial neural-network models. They conclude that the increased data richness has led to improved modeling techniques; however, they suggest that future work will need to develop novel techniques to incorporate this information and ultimately elucidate water consumption relationships at multiple scales. Furthermore, they identify three key characteristics that may lead to a generally-accepted water consumption modeling framework:

- Development of water consumption models for practical application should focus on those models with input variables that can be easily collected, monitored, and used by the utility.
- Water consumption models should be as parsimonious as possible without compromising the integrity of their forecasting quality.
- Future development should focus on probabilistic forecasting methods that allow utilities to make decisions, while quantifying the level of uncertainty of the resulting water consumption forecasts.

These authors acknowledge that there is no clear answer to the question: "Which model is best for water consumption forecasting?" and state that current water consumption modeling applications require specific parameterization and implementation for different geographic locations, water rate structures, historical data quality, and periodicity. Therefore, there is a need in this industry to

develop a general water consumption forecasting approach that broadly applies by overcoming the specificity of the current models and methodologies. The objective of this analysis is to provide a methodology for quantifying the influence of ambient processes on the water consumption distribution discussed by Donkor et al. This distribution comprises the water consumption measurements from all residential accounts and allows the estimation of the arithmetic mean water consumption. Viewing consumption as a distribution may allow the water utility to better understand how the collective residential account holders are changing their consumption habits. Ambient processes include price, temperature, precipitation, as well as water conservation, education, and by-law enforcement.

The foundation of this analysis is the set of residential water consumption data collected from the City of Waterloo, Ontario, Canada. Specifically, this data comprises water consumption meter readings representing 10 years consisting of 60 bimonthly billing periods between January/February 2007 and November/December 2016 for a total of 1,549,371 observations and 51,291,348 m^3 of cumulative billed water. The pricing structure of the water utility is a volume-constant rate and the utility services upwards of 27,000 residential accounts during each billing period. Summary data is provided in Appendix B.1. Figure 3.1 provides PMF data from the November/December bimonthly billing period for the years 2008, 2010, 2012 and 2014. PMF data is computed using a frequency histogram that counts the number of consumers in 1 m^3 water consumption bins and subsequently divides this number by the total measurements within a sampling interval. The utility has been annually increasing its real water price, with nearly a one-real dollar per cubic meter (40-percent) increase between 2008 and 2014. This annual price increase coincides with declining water consumption characterized by a progressive compression of the water consumption PMF toward the origin. The smooth unimodal shape of these PMFs may indicate that residential water consumption is a continuum process, where the location, scale and shape is subject to ambient influences such as price and weather.

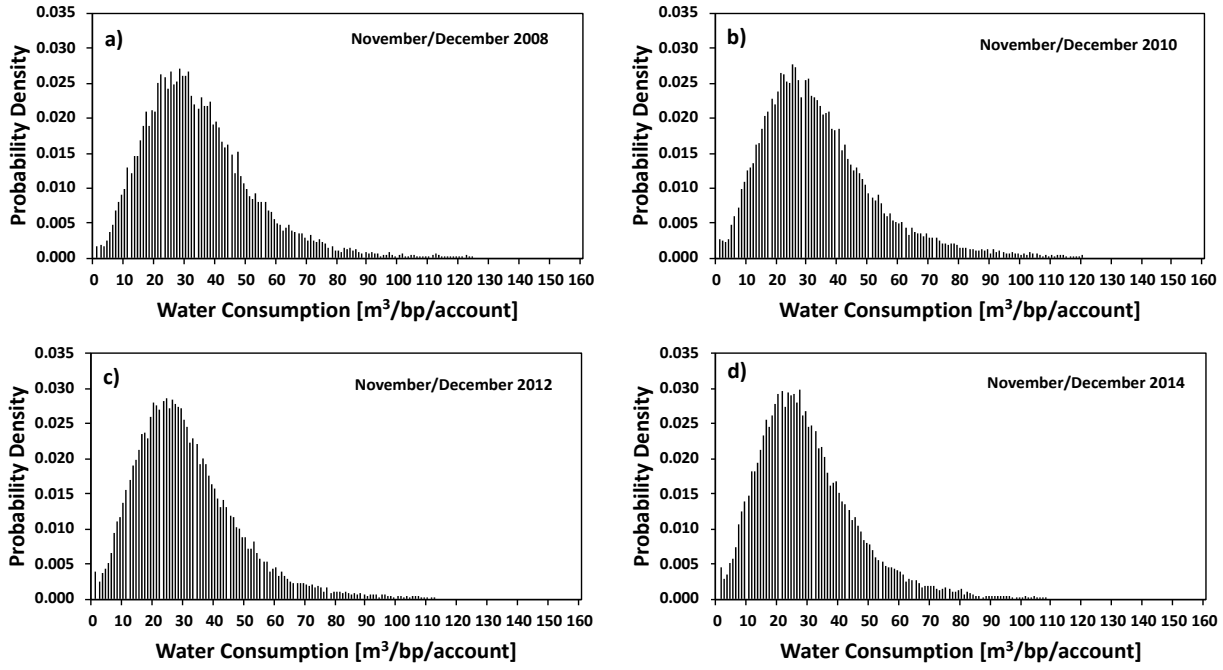

Figure 3.1: Select residential water consumption PMFs for November/December from the City of Waterloo, Ontario, Canada. Figure 3.1.a presents data from 2008; Figure 3.1.b presents data from 2010; Figure 3.1.c presents data from 2012; and Figure 3.1.d presents data from 2014. Note that weather conditions are generally consistent during the November/December billing periods presented on Figure 3.1.

The methodology begins by transforming all of the discrete histograms for each of the 60 bimonthly billing periods into smooth and continuously differentiable PDFs as outlined in Chapter 2. This step involves measuring discrete statistics for each sampling period, and then choosing a representative but normalized parametric PDF p_z in the standard-score space which has a consistent form throughout the analysis. All sampling periods are characterized by a unimodal PDF that is shifted and asymmetric with a heavy tail. The median of the dataset m_x is used to measure the location of the water consumption PMF, the standard deviation σ_x measures its scale, while the control function parametrization for the standard-score PDF p_z characterizes its shape.

This methodology closely resembles that of kernel density estimation techniques for econometrics applications (Zambom and Dias 2012) with the following important differences. Zambom and Dias identify some drawbacks of kernel density estimation techniques and state that bounded data always produces a biased estimate near the data boundaries. For instance, spurious noise can occur within the tail of an estimation or important features of the distribution may be

lost due to over-smoothing when the underlying density has a characteristic long tail. The methodology applied to water consumption data herein addresses these shortcomings to reproduce long-tailed distributions as continuous functions that evolve through time. Strategic data culling paired with the objective function proposed in chapter 2 are designed to limit estimator bias, without sacrificing the ability to accurately reproduce the long-tailed nature of some datasets when fitting a parametric PDF. The consistent parametrization obtained via the control function theory is critical to assess how the shape of the water consumption PDF evolves through time as a continuous process.

The proposed water consumption framework involves formulating and parameterizing a model that has the mathematical structure of an "advection-dispersion" solution. This formulation results in a transport model to propagate the time-sequence of PDFs representing the continuum response of the residential water consumption to the ubiquitous, measurable, and regular (if not continuous) processes of price P and weather W. Omitted ambient processes may include but are not limited to passive water conservation, education, and by-law enforcement. To begin, a model is proposed using the total differential of a set of statistics that inform the advective-dispersive process. These statistics include the median, standard deviation, control function parameterization as attributes that characterize both advection and dispersion. The functional relationship of each statistic is obtained by expanding the total differential with respect to ambient processes. This approach results in a formulation that is consistent with curvilinear regression. Regression serves as the basis for parametrizing a time-continuous advective-dispersion model using discrete observations of statistics derived by optimally fitting PDFs to the histogram and PMF data, as well as real water price and temperature/precipitation data, from each sequential sampling period.

Knowledge of how the mean statistic changes as a function of ambient processes such as price and weather is critical for the water utility to forecast future water consumption for financial sustainability. The mean statistic is a scalar measure that represents the magnitude of any continuum response observed in the water consumption PMF. In the case of the water utility, the forecasted mean statistic multiplied by the number of active accounts is the expected value of total water consumption. The proposed framework provides two independent estimates of the mean statistic. The first is derived by integrating the PDF from the advective-dispersion transport model and is denoted here as the "transport mean". The second is obtained by application of a curvilinear regression model to the discrete arithmetic data as a function of price P and weather W and is

denoted as the "direct regression mean". The key difference is that the direct regression mean is independent of the statistics that define the water consumption data as a PMF such as the median, standard deviation, and control function. An outcome of this analysis is to show that the transport mean provides unique information relative to the direct regression mean. The transport model performs regression on each of the median, standard deviation, and control function parameters and combines these models to reproduce the advective-dispersive characteristics of the water consumption PMF. The transport model then indirectly evaluates the resulting mean statistic.

A potential benefit of the proposed water consumption framework is that the influence of unmeasurable ambient processes can be observed in the PDF as a solution to the transport model. This analysis will show that implementation of that water conservation policy in the last two years for which data was available for this analysis created a significant departure in the location, scale and shape of the PDF relative to that which the advective-dispersive transport model could provide. However, the PMF data could be accurately fit with the control function approach implying that it reflected a continuum response of consumer water consumption behaviour. This work suggests that future inclusion of a "water conservation policy" as a quantified ambient process parameter within the advective-dispersive transport framework could remedy this issue.

3.1. MODEL DEVELOPMENT

Advective-dispersive transport is used to model the temporal evolution of utility-wide residential water consumption distribution as a function of changes in ambient process, such as the unit price of water and weather conditions. The water consumption response at any observation interval is represented by a histogram, which is then transformed into a PMF and finally a PDF which allows the mean statistic to be determined. Consequently, the temporal evolution of the PDF and mean statistic represent the solution to an advective-dispersive transport problem for a continuum system. Here, the continuum system represents the utility-wide residential water consumption. Chapter 2 develops the control function theory for quantifying the location, scale, and shape of a measurement space PDF p_x using the median m_x, standard deviation σ_x, and standard-score PDF p_z, respectively. In this context, the median represents the bulk translation (advection) of the distribution, while the standard deviation and standard-score PDF combine to characterize the relative frequency or spread (dispersion) of the data as:

$$\underbrace{p_x}_{\substack{continuum\\distribution}} = \underbrace{m_x}_{\substack{probabilistic\\advective\\process}} + \underbrace{\frac{1}{\sigma_x} \times p_z}_{\substack{probabilistic\\dispersive\\process}} \tag{3.1}$$

Chapter 2 also demonstrates that the mean statistic of the measurement space data μ_x is a scalar value that describes the ensemble magnitude of the measurement space PDF p_x as a continuum system. Using this premise, the mean statistic also represents an advective-dispersive process as:

$$\mu_x = \underbrace{\underbrace{m_x}_{\substack{distribution\\location}}}_{\substack{mean\\advective\\process}} + \underbrace{\underbrace{\sigma_x}_{\substack{distribution\\scale}} \times \underbrace{\mu_z}_{\substack{distribution\\symmetry}}}_{\substack{mean\\dispersive\\process}} \tag{3.2}$$

Where the standard-score mean μ_z quantifies the symmetry of the distribution and has a value of $\mu_z = 0$ for symmetric distributions. The symmetry of the distribution is solely dependent upon the definite integral of the position-weighted standard-score PDF p_z and the control function parameters that describe it. Given the conditional dependence of μ_z on p_z in the context of advective-dispersive transport in Equation 3.2, we can present the mean statistic as a projection of the PDF symmetry through its location and scale as: $\mu_z = \int z p_z \, dz$. Notably, a distribution has perfect symmetry for values of $\mu_z = 0$ with this value growing as the distribution becomes increasingly asymmetric. Perfect symmetry results in the advective-dispersive process of the mean in Equation 3.2 being solely dependent upon the median m_x.

Development of the advection-dispersion model for p_x and μ_x in the sections below begins by evaluating the discrete statistics of the raw data from each sampling period and then transforming them into a time-continuous form. This results in a time-continuous PDF whose parametric values can be adjusted so as to be able to reproduce the entire sampling sequence of discrete histogram information. The parametric values allow the PDF to change location, scale and shape in response to a set of continuum ambient processes, which change as a function of time t over sequential sampling intervals. These ambient processes are denoted using the variable "$\mathbb{x}_j$". Note that the following model development remains general with respect to the relationship between the statistics for the median m_x, standard deviation σ_x, as well as the standard-score PDF p_z and the ambient processes $\mathbb{x}_j(t)$.

3.1.1. DISCRETE STATISTICS

The discrete statistics are scalar measures of the location $m_{x,i,t}$, scale $\sigma_{x,i,t}$, and mean statistic $\mu_{x,i,t}$ of a "continuum system" because they reflect the aggregation of all of the active accounts within a sampling period t into a single distribution. Notably, t represents the sampling interval of the analysis described by each discrete statistic. The arithmetic mean characterizes the magnitude of the discrete data for some sampling interval t and can be expressed as follows:

$$\mu_{x,i,t} = \frac{1}{N_{i,t}} \sum_{i=1}^{N_i} x_{i,t}\,, \qquad \mu_{z,i,t} = \frac{\mu_{x,i,t} - m_{x,i,t}}{\sigma_{x,i,t}} \tag{3.3}$$

Where, water consumption measured for any account i in the discrete sampling interval (billing period) t is denoted as $x_{i,t}$ and $N_{i,t}$ represents the number of active residential accounts within the utility at each sampling interval. Knowledge of the discrete median $m_{x,i,t}$ and standard deviation $\sigma_{x,i,t}$ allow for transformation of the mean $\mu_{x,i,t}$ into a representation of distribution symmetry $\mu_{z,i,t}$ in the standard-score space. The discrete median $m_{x,i,t}$ and standard deviation $\sigma_{x,i,t}$ are quantified for each billing period as:

$$\begin{aligned} m_{x,i,t} &= \left\{\frac{N_{i,t}+1}{2}\right\}^{th} \text{ value} \\ \sigma_{x,i,t} &= \sqrt{\frac{1}{(N_{i,t}-1)} \sum_{i=1}^{N_{i,t}} \left[x_{i,t} - m_{x,i,t}\right]^2} \end{aligned} \tag{3.4}$$

Chapter 2 provides a thorough explanation of how the discrete median $m_{x,i,t}$ and standard deviation $\sigma_{x,i,t}$ relate to the statistical transformations of the mean statistic between the measurement space x, median-relative space y, and standard-score space z.

Upon evaluating the discrete statistics for each sampling interval, the analysis can hypothesize a formulation of the control function that will be adequate for reproducing the shape of the discrete histogram. This analysis applies a third-order exponential polynomial control function developed in chapter 2 as:

$$p_{z,t} = \exp\left(-\int g_{z,t}\, dz\right), \qquad g_{z,t} = \left[\alpha_{1,t} + \alpha_{2,t} z + \alpha_{3,t} z^2 + \alpha_{4,t} z^3\right] \tag{3.5}$$

Where, $z = \frac{x - m_{x,i,t}}{\sigma_{x,i,t}}$ is the standard-score transformation from the measurement space x; $p_{z,t}$ represents the continuous standard-score PDF; $g_{z,t}$ represents the control function; and, $\alpha_{n_z,t}$ represent the control function parameters for each histogram at sampling interval t. Notably, the standard-score PDF $p_{z,t}$ for each sampling interval transforms into the measurement space PDF $p_{x,t}$ using the discrete median $m_{x,i,t}$ and standard deviation $\sigma_{x,i,t}$ using Equation 3.1 as: $p_{x,t} = m_{x,i,t} + \frac{1}{\sigma_{x,i,t}}\, p_{z,t}$.

The probability weighted mean $\mu_{z,t}$ or symmetry of the distribution in the standard score space z is derived from the parametric PDFs $p_{z,t}$ at each sampling interval t as:

$$\mu_{z,t} = \int_{z_{min}}^{z_{max}} z p_{z,t}\, dz \tag{3.6}$$

Where, $z_{min} = \frac{x_{min} - m_{x,i,t}}{\sigma_{x,i,t}}$ and $x_{min} = 0$; and $z_{max} = \frac{x_{max} - m_{x,i,t}}{\sigma_{x,i,t}}$ and $x_{max} = 4m_{x,i,t}$ from data culling. Equation 3.6 shows a relationship between symmetry $\mu_{z,t}$ and the control function $g_{z,t}$ through the standard-score PDF $p_{z,t}$ and is the discrete representation of the symmetry estimator for data within each sampling interval. Evaluating the discrete standard-score mean $\mu_{z,t}$ for the corresponding PDF during each sampling interval t requires numerical integration of Equation 3.6. Finally, $\mu_{z,t}$ can be transformed from the standard-score space z into the measurement space x as $\mu_{x,t}$ using the discrete median $m_{x,i,t}$ and standard deviation $\sigma_{x,i,t}$ from Equation 3.2 as: $\mu_{x,t} = m_{x,i,t} + \sigma_{x,i,t}\, \mu_{z,t}$.

3.1.2. TIME-CONTINUOUS STATISTICS

Time-continuous statistics are an extension of the above discrete statistics in that they are a continuous function of an ambient process as well as time. The process of defining the continuous statistics begins with the total differential with respect to two independent variables $\mathbb{x}_1$ and $\mathbb{x}_2$ using placeholder variable $U \in \{\mu_x, m_x, \sigma_x, \alpha_{n_z}\}$. Notably, $\mathbb{x}_1$ and $\mathbb{x}_2$ represent the relevant ambient processes of interest. The total differential of U can be expressed as:

$$dU \equiv F'_U d\mathbb{x}_1 + F^*_U d\mathbb{x}_2 \tag{3.7}$$

Where, $\alpha_{n_z} \in \{\alpha_1 \dots \alpha_4\}$ for a third-order control function. F'_U represents the partial derivative of U with respect to $\mathbb{x}_1$ as $F'_U\big(\mathbb{x}_1(t), \mathbb{x}_2(t)\big) \equiv \frac{\partial U}{\partial \mathbb{x}_1}$. Similarly, F^*_U represents the partial derivative of U with respect to $\mathbb{x}_2$ as $F^*_U\big(\mathbb{x}_1(t), \mathbb{x}_2(t)\big) \equiv \frac{\partial U}{\partial \mathbb{x}_2}$. Note that "$(t)$" denotes a time-dependent process. Then it follows that the statistics describing a PDF can inherit their time-dependence from these processes: $U\big(\mathbb{x}_1(t), \mathbb{x}_2(t)\big)$. Therefore, with knowledge of the time-ordered nature of $\mathbb{x}_1$ and $\mathbb{x}_2$, the influence of these ambient processes can be projected onto the progression of the statistics that describe the distribution and its mean. This results in an advective-dispersive representation of a PDF as it evolves through time from Equation 3.1. Similarly, Equation 3.2 describes the advective-dispersive process controlling the evolution of the mean statistic. Equation 3.7 presents a model where $\mathbb{x}_1$ and $\mathbb{x}_2$ will correlate with the continuum statistics that describe the discrete histogram. What remains is a formal expansion of F'_U and F^*_U for each statistic with respect to each independent variable. Expand F'_U and F^*_U using a Taylor expansion around the point $\mathbb{x}_1 = 0$ and $\mathbb{x}_2 = 0$ as:

$$\begin{aligned}
F'_U - F'_{U,0} &= \frac{\partial F'_U}{\partial \mathbb{x}_2} d\mathbb{x}_2 + \frac{\partial F'_U}{\partial \mathbb{x}_1} d\mathbb{x}_1 + 2\frac{\partial^2 F'_U}{\partial \mathbb{x}_1 \partial \mathbb{x}_2} d\mathbb{x}_2 d\mathbb{x}_1 + \frac{1}{2}\frac{\partial^2 F'_U}{\partial \mathbb{x}_1^2} d\mathbb{x}_1^2 \\
&\quad + \frac{1}{2}\frac{\partial^3 F'_U}{\partial \mathbb{x}_1^2 \partial \mathbb{x}_2} d\mathbb{x}_2 d\mathbb{x}_1^2 + \cdots \\
F^*_U - F^*_{U,0} &= \frac{\partial F^*_U}{\partial \mathbb{x}_1} d\mathbb{x}_1 + \frac{\partial F^*_U}{\partial \mathbb{x}_2} d\mathbb{x}_2 + 2\frac{\partial^2 F^*_U}{\partial \mathbb{x}_1 \partial \mathbb{x}_2} d\mathbb{x}_2 d\mathbb{x}_1 + \frac{1}{2}\frac{\partial^2 F^*_U}{\partial \mathbb{x}_2^2} d\mathbb{x}_2^2 \\
&\quad + \frac{1}{2}\frac{\partial^3 F'_U}{\partial \mathbb{x}_2^2 \partial \mathbb{x}_1} d\mathbb{x}_1 d\mathbb{x}_2^2 + \cdots
\end{aligned} \tag{3.8}$$

Then, substitute and integrate F'_U and F^*_U within the total differential to generate a curvilinear regression model for $U\big(\mathbb{x}_1(t), \mathbb{x}_2(t)\big)$. To compress notation let $F''_U = \frac{\partial F'_U}{\partial \mathbb{x}_1}$, $F'''_U = \frac{\partial^2 F'_U}{\partial \mathbb{x}_1^2}$; $F^{**}_U = \frac{\partial F^*_U}{\partial \mathbb{x}_2}$, $F^{***}_U = \frac{\partial^2 F^*_U}{\partial \mathbb{x}_2^2}$; $F^{*\prime}_U = \frac{\partial F'_U}{\partial \mathbb{x}_2}$, $F^{*\prime\prime}_U = \frac{\partial^2 F'_U}{\partial \mathbb{x}_1 \partial \mathbb{x}_2}$; $F^{*\prime}_U = \frac{\partial F^*_U}{\partial \mathbb{x}_1}$, $F^{**\prime}_U = \frac{\partial^2 F^*_U}{\partial \mathbb{x}_1 \partial \mathbb{x}_2}$; and, $F^{*\prime\prime\prime}_U = \frac{\partial^3 F'_U}{\partial \mathbb{x}_1^2 \partial \mathbb{x}_2}$ and $F^{***\prime}_U = \frac{\partial^3 F^*_U}{\partial \mathbb{x}_2^2 \partial \mathbb{x}_1}$. Substitution into Equation 3.9 produces a general relationship that solves for dU as:

$$dU \equiv \Big[F'_{U,0} + F^{*\prime}_{U} d\mathbb{x}_2 + F''_{U} d\mathbb{x}_1 + 2F^{*\prime\prime}_{U} d\mathbb{x}_2 d\mathbb{x}_1 + \frac{1}{2}F'''_{U} d\mathbb{x}_1^2 + \frac{1}{2}F^{*\prime\prime\prime}_{U} d\mathbb{x}_2 d\mathbb{x}_1^2 + \cdots\Big] d\mathbb{x}_1 + \Big[F^{*}_{U,0} + F^{*\prime}_{U} d\mathbb{x}_1 + F^{**}_{U} d\mathbb{x}_2 + 2F^{**\prime}_{U} d\mathbb{x}_2 d\mathbb{x}_1 + \frac{1}{2}F^{***}_{U} d\mathbb{x}_2^2 + \frac{1}{2}F^{***\prime}_{U} d\mathbb{x}_1 d\mathbb{x}_2^2 + \cdots\Big] d\mathbb{x}_2 \tag{3.9}$$

This expression for dU can be used to forecast how the mean statistic μ_x changes as a function of $\mathbb{x}_1$and $\mathbb{x}_2$. Moreover, dU can be adapted into the form of a transport model to forecast the median m_x, standard deviation σ_x, and control function parameters $\alpha_{n_z,t}$. The intent of the transport model is to reproduce the trends of the entire PDF and indirectly evaluate the mean statistic as a function of ambient processes $\mathbb{x}_1$and $\mathbb{x}_2$.

3.1.3. ADVECTIVE-DISPERSIVE TRANSPORT WITH AMBIENT PROCESSES

To proceed with adapting Equation 3.11 for curvilinear regression, first truncate the terms in Equation 3.9 at $F'''_U = F^{***}_U = F^{***\prime}_U = F^{*\prime\prime\prime}_U = 0$ to an approximate value $dU \cong d\widehat{U}$. For the condition that $d\mathbb{x}_1 = \mathbb{x}_1 - 0$ and $d\mathbb{x}_2 = \mathbb{x}_2 - 0$ the relationship for $\widehat{U}$ simplifies as:

$$\int_{\widehat{U}_0}^{\widehat{U}} d\widehat{U} = F'_{U,0}\mathbb{x}_1 + F^{*}_{U}\mathbb{x}_2 + 2F^{*\prime}_{U}\mathbb{x}_2\mathbb{x}_1 + \frac{1}{2}F^{**}_{U}\mathbb{x}_2^2 + F^{**\prime}_{U}\mathbb{x}_2^2\mathbb{x}_1 + \frac{1}{2}F''_{U}\mathbb{x}_1^2 + F^{*\prime\prime}_{U}\mathbb{x}_2\mathbb{x}_1^2 \tag{3.10}$$

Next, all partial derivatives in Equation 3.10 are expressed as coefficients $b_{U,1\ldots7}$ to succinctly express $\widehat{U}$ as a curvilinear regression model:

$$\widehat{U} = b_{U,0} + b_{U,1}\mathbb{x}_1 + b_{U,2}\mathbb{x}_2 + b_{U,3}\mathbb{x}_2\mathbb{x}_1 + b_{U,4}\mathbb{x}_2^2 + b_{U,5}\mathbb{x}_2^2\mathbb{x}_1 + b_{U,6}\mathbb{x}_1^2 + b_{U,7}\mathbb{x}_2\mathbb{x}_1^2 \tag{3.11}$$

where, $\widehat{U}_0 = b_{U,0}$ and $b_{U,0\ldots7}$ represent the curvilinear regression parameters.

To proceed with a general representation of Equation 3.1 as an advective-dispersive transport model under the influence of ambient processes, first condense the notation in Equation 3.11 with $\widehat{U} \in \{m_x, \sigma_x, \alpha_{1\ldots4}\}$. Next, re-express Equation 3.1 as:

$$\hat{p}_x = \hat{m}_x + \frac{1}{\hat{\sigma}_x} \exp\left(-\int [\hat{\alpha}_1 + \hat{\alpha}_2 z + \hat{\alpha}_3 z^2 + \hat{\alpha}_4 z^3]\, dz\right) \tag{3.12}$$

Similarly, advective-dispersive transport of the mean statistic following Equation 3.2 is expressed as:

$$\hat{\mu}_{x|\hat{p}_x} = \hat{m}_x + \hat{\sigma}_x \int_{z_{min}}^{z_{max}} z\ \exp\left(-\int [\hat{\alpha}_1 + \hat{\alpha}_2 z + \hat{\alpha}_3 z^2 + \hat{\alpha}_4 z^3]\, dz\right) dz \tag{3.13}$$

where z_{min} and z_{max} is the range of the integration in accordance with Equation 3.6. Both Equations 3.12 and 3.13 indicate that advective-dispersive transport occurs as the location, scale, and shape of the continuum distribution of observed measurements respond to ambient processes. Note the subscript notation "$|\hat{p}_x$" is used to differentiate the transport model mean from the direct regression mean $\hat{\mu}_x$.

Validation of advective-dispersive transport process for the mean statistic follows by using Equation 3.11 to directly regress the response of the mean statistic to ambient process. Notation for this process is given as:

$$\hat{\mu}_x = b_{\mu,0} + b_{\mu,1}\mathbb{x}_1 + b_{\mu,2}\mathbb{x}_2 + b_{\mu,3}\mathbb{x}_2\mathbb{x}_1 + b_{\mu,4}\mathbb{x}_2^2 + b_{\mu,5}\mathbb{x}_2^2\mathbb{x}_1 + b_{\mu,6}\mathbb{x}_1^2 + b_{\mu,7}\,\mathbb{x}_2\mathbb{x}_1^2 \tag{3.14}$$

Equation 3.13 is referred to as the “transport model mean”, while Equation 3.16 is referred to as the “direct regression mean”. Equivalence of 3.13 and 3.14 implies that the magnitude of the ensemble continuum response to ambient process can be inferred without knowledge of the location, scale and shape of the distribution of observations itself. However, this information obviously exists and serves to constrain the range of measurable data constituting the continuum response, as represented by the PMF. Moreover, the transport model $\hat{p}_x$ guarantees a unique solution for the mean statistic $\hat{\mu}_{x|\hat{p}_x}$ by replicating the PMF of the raw data. This is in contrast to the direct regression mean $\hat{\mu}_x$ which does not constrain combinations of the location, scale, and shape of the distribution. Therefore, working with the direct regression mean $\hat{\mu}_x$ ignores the availability of the location, scale, and shape information describing the PMF. This is important because there are infinite combinations of location, scale, and shape that can generate the same mean statistic for any continuum system. In contrast, the transport model guarantees a unique mean statistic solution for any set of ambient conditions.

3.2. RESIDENTIAL WATER CONSUMPTION APPLICATION

Empirical evidence for the continuum response of utility-wide residential water consumption is presented in Figure 3.1 where the PMFs of measurement data changed location, scale and shape due to increases in the real unit price of water. The previous section develops an advective-dispersive transport model to quantify how the PDF $\hat{p}_x$ represents the continuum response of residential water consumption to ambient processes such as price. In addition to real water price P, it is anticipated that weather W as a representation of temperature and precipitation, water restriction by-law enforcement, water conservation, and education are key ambient processes impacting water consumption. While P and W are observable ambient processes that can be measured and recorded, public policy initiatives such as by-law enforcement and education are difficult to quantify in the same manner. However, the response of the water consumption histogram to changes in P, W is tangible. It is expected that the impact of policy and education on water consumption can only be inferred via observed changes in the water consumption PDF beyond those that can be explained via tangible ambient processes. The development of the continuous statistics above culminating in Equation 3.13 shows that the advective-dispersive transport of $\hat{p}_x$ is dependent on knowledge of the median, standard deviation, and control function statistics of the water consumption dataset. Appendix B.1 provides the discrete statistics $m_{x,i,t}$, $\sigma_{x,i,t}$, and $\mu_{x,i,t}$ for each sampling interval t within the 60 bimonthly periods considered. Here, this analysis denotes the set of these continuum ambient processes that change during each sampling period t using the variable $\mathbb{x}_j \in \{P(t), W(t)\}$.

This analysis proceeds in three steps. First, it applies the methodology from Chapter 2 to transform water consumption histograms (see Figure 3.1) into optimally parameterized and continuous PDFs that are consistent with the advective-dispersive processes expressed in Equations 3.1 and 3.2. Second, the analysis performs curvilinear regression upon the median, standard deviation, and control function parameters with real water price and weather score. Statistically defensible correlation to ambient processes supports the contention that the water consumption PMF represents a continuum system that experiences advective-dispersive transport. Third, the analysis compares the direct regression model to the transport model estimates for the mean statistic. This section builds experimental evidence using the above advection-dispersive transport theory to justify the empirical observation that the water consumption PMFs exhibit a

continuum response to ambient processes. The outcome of the following analysis is that the transport model is at least as effective as the direct regression model for estimating the evolution of the mean statistic.

3.2.1. PARAMETRIC PDFS AS A REPRESENTATION OF THE CONTINUUM RESPONSE

Applying the control function theory from Chapter 2 to the PMF data shown on Figure 3.1 produces optimally parameterized continuous functions that compress each sampling period t dataset into a median $m_{x,i,t}$, standard deviation $\sigma_{x,i,t}$, and control function parameters $\alpha_{1,t}$, $\alpha_{2,t}$, $\alpha_{3,t}$, and $\alpha_{4,t}$. The discrete statistics $m_{x,i,t}$ and $\sigma_{x,i,t}$ provide scalar estimates of the location and scale of the observation data $x_{i,t}$ while the shape of the CMF is captured by the "best fit" parametric PDF $p_{z,t}$ (see Equation 3.5) through the control function $g_{z,t}$. These statistics are estimated by matching cumulative distribution functions (CDFs) to the CMFs derived from the culled data. Appendix B.2 lists all control function parameters for each bimonthly period t between January/February 2007 through November/December 2016. The resulting CMFs are listed in Appendix B.3. As previously mentioned, this parametrization proceeds on the basis that limited data culling of the observation data $x_{i,t}$ is necessary to remove measurements that are greater than four-times the median $m_{x,i,t}$ of each sampling period t. The remaining data reflect greater than 98-percent of the original data for all billing periods considered herein. Water consumption measurements beyond this threshold include multi-unit dwellings and extreme residential water consumers, which do not reflect the water consumption behaviour for the population of interest in this analysis.

The control function parameterization is adjusted to ensure that the shape of the CDF matches that of the CMF for each sampling period. This proceeds by minimizing the objective function shown in Equation 2.10. This process relies on the hierarchal relationship between the control function, PDF, and CDF to ensure that $p_{x,t}$ correctly reproduces the PMF over the entire range of observation data $x_{i,t}$. This step provides evidence that $p_{x,t}$ reproduces the continuum process represented by the PMF for sampling period t, and that $\mu_{x,t}$ is a unique representation of $\mu_{x,i,t}$. Table 3.1 presents the mean square error estimates $MSE = \sqrt{\left[\mu_{x,t} - \mu_{x,i,t}\right]^2}$ for each sampling interval to quantify the departure of the continuous distribution from the raw data. MSE values of the mean water consumption are always less than $1\,m^3/bp/acct$ which is the

measurement accuracy for each meter reading $x_{i,t}$, where "bp" denotes a billing period and "$acct$" denotes account. Figure 3.2 presents the PDF with the optimal parameterization for the July/August billing period during 2007, 2009, 2015 and 2016 superimposed onto its respective PMF to demonstrate the goodness of fit over the entire range of observation data $x_{i,t}$. To varying degrees, each water consumption PMF is reproduced by PDFs that are asymmetric, shifted, and exhibit a heavy tail.

Table 3.1: MSE values between observed and optimally parameterized mean statistic. Values represents MSE between arithmetic mean water consumption $\mu_{x,i,t}$ and $\mu_{x,t}$ estimated from the fitted PDF.

	$MSE\ [m^3/bp/account]$					
Year	**Jan/Feb**	**Mar/Apr**	**May/June**	**July/Aug**	**Sept/Oct**	**Nov/Dec**
2007	5.14×10^{-3}	9.00×10^{-8}	1.28×10^{-1}	1.00×10^{-1}	1.15×10^{-1}	1.02×10^{-3}
2008	4.76×10^{-4}	1.14×10^{-1}	1.38×10^{-1}	8.93×10^{-3}	1.09×10^{-1}	1.83×10^{-3}
2009	4.00×10^{-3}	1.45×10^{-5}	1.06×10^{-3}	5.22×10^{-3}	1.03×10^{-2}	6.74×10^{-3}
2010	2.63×10^{-3}	4.48×10^{-4}	2.67×10^{-2}	2.54×10^{-2}	1.23×10^{-2}	1.61×10^{-2}
2011	1.70×10^{-2}	5.21×10^{-3}	1.99×10^{-2}	1.47×10^{-1}	5.01×10^{-4}	5.81×10^{-3}
2012	5.10×10^{-3}	1.70×10^{-1}	2.04×10^{-1}	1.21×10^{-2}	1.94×10^{-4}	1.49×10^{-1}
2013	2.96×10^{-3}	1.53×10^{-1}	8.42×10^{-3}	3.80×10^{-3}	1.52×10^{-1}	1.95×10^{-1}
2014	1.41×10^{-1}	4.26×10^{-4}	1.40×10^{-3}	2.17×10^{-1}	1.03×10^{-3}	1.21×10^{-4}
2015	7.96×10^{-1}	5.12×10^{-1}	4.39×10^{-3}	9.28×10^{-2}	3.89×10^{-1}	3.89×10^{-1}
2016	1.09×10^{-2}	2.48×10^{-2}	5.16×10^{-2}	4.34×10^{-2}	4.73×10^{-2}	6.08×10^{-3}

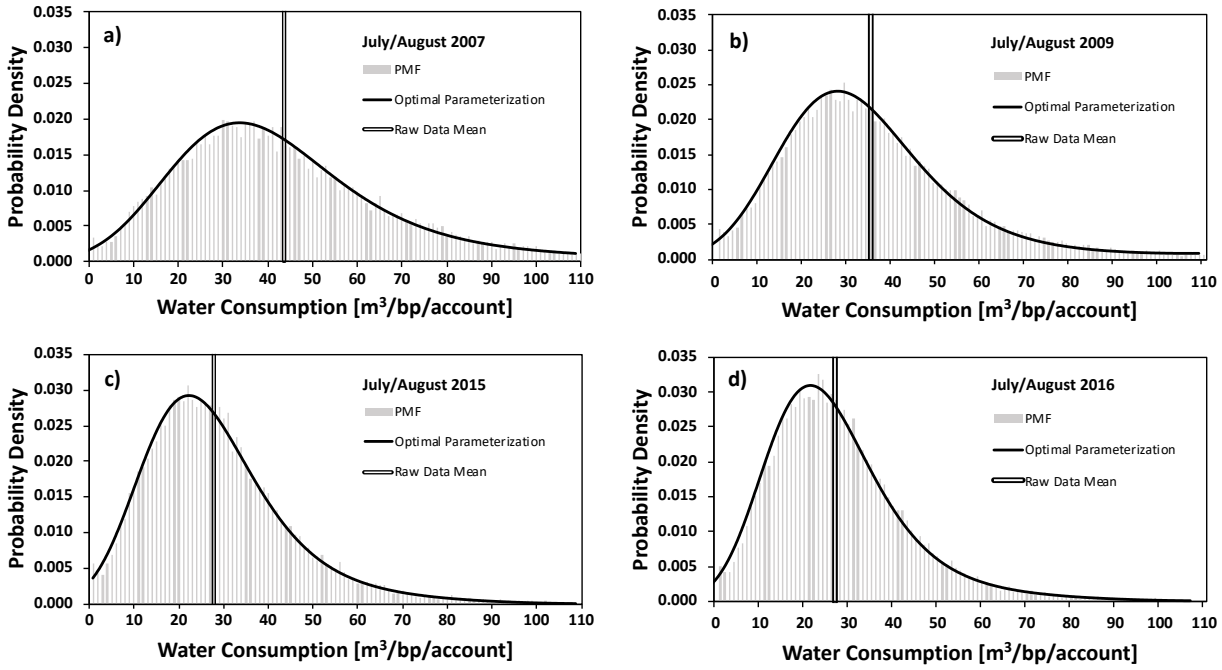

Figure 3.2: Select water consumption data and optimal parametric fit. Residential water consumption PMFs and their corresponding best fit PDFs from July/August billing period in 3.2.a) 2007, 3.2.b) 2009, 3.2.c) 2015, and 3.2.d) 2016. Also shown is the arithmetic mean $\mu_{x,i,t}$ of the raw data.

Figure 3.3 qualitatively illustrates the impact of real water price on water consumption. Specifically, water consumption PMFs and corresponding best fit parametric PDFs are shown that span biannual increments of the November/December bimonthly periods from 2008, 2010, 2012 to 2014. During this bimonthly period, outdoor water usage in Waterloo, Ontario, Canada is minimal due to the onset of cold winter weather and dormant vegetation. Therefore, it is assumed that variations in outdoor temperature and precipitation during this bimonthly period do not meaningfully impact residential water consumption. However, annual increases in the real water price over this eight-year period do impact the location, scale and shape of the water consumption PDF. Progressive increases in the real price of water have caused the water consumption distribution to compress toward the origin. This compression includes a shift of the mode as well as a reduction in the length and extent of the tail, thereby illustrating a continuum response of the water consumption PDF $p_{x,t}$. Moreover, as the entire distribution or continuum shifts with price increases, then the mean statistic also changes with respect to price. As expected, the mean statistic $\mu_{x,t}$ for the November/December period decreases from 34.29, 33.76, 31.06, to 29.71

$[m^3/bp/acct]$ during 2008, 2010, 2012 to 2014, respectively. These results indicate that increasing the real water price negatively influences residential water consumption. It is worthwhile to note that there are many confounding variables that also negatively influence water consumption beyond real water price increases. These include: education, passive conservation, technological advances in water efficiency, and water policy. However, it is suspected that these influences are minimal when compared to the economic disincentive provided by real water price increases.

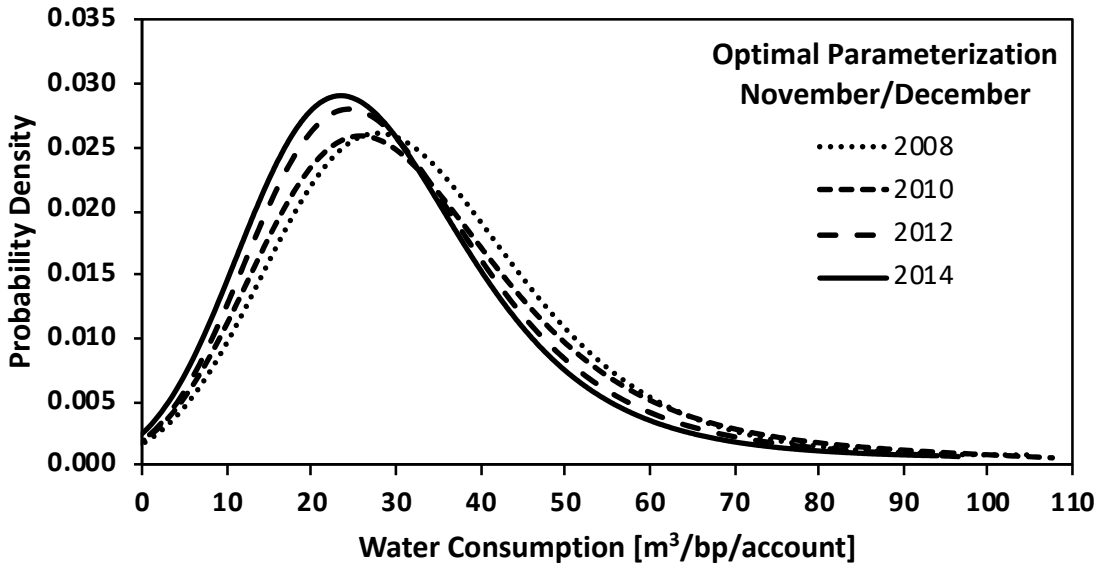

Figure 3.3: Select optimally parameterized PDFs for November/December billing period.
Continuum response of the water consumption PDF $p_{x,t}$ to real water price using select parametric PDFs from the November/December billing period.

Figure 3.4 demonstrates the co-dependence of consumer water consumption behaviour to changing weather conditions as well as the real water price. Optimally fit PDFs $p_{x,t}$ for three bimonthly billing periods in 2007, 2009, 2015, and 2016 are presented. Note that the real price of water is constant throughout 2007, and progressively increases for each of 2009, 2015, and 2016. Waterloo experiences summer weather conditions between July and October each year that are characterized by warm to hot temperatures and intermittent precipitation, causing residents to have their greatest demand for outdoor water use, especially for irrigation. However,

September/October represents early fall and the end of the growing season with declining need for irrigation. Finally, January/February represents winter conditions with below-freezing temperatures where residents have limited need for outdoor water usage. Figure 3.4 shows that the water consumption PDF compresses toward the origin during the winter months and expands out from the origin during the summer months. However, consumers appear to be more sensitive to the influence of weather in 2007 than 2016, with their sensitivity to weather gradually decreasing through time as the real water price increases. In fact, water consumption behaviour appears seasonally stagnant in 2016. This observation qualitatively supports the above hypothesis that both price and weather influence water consumption.

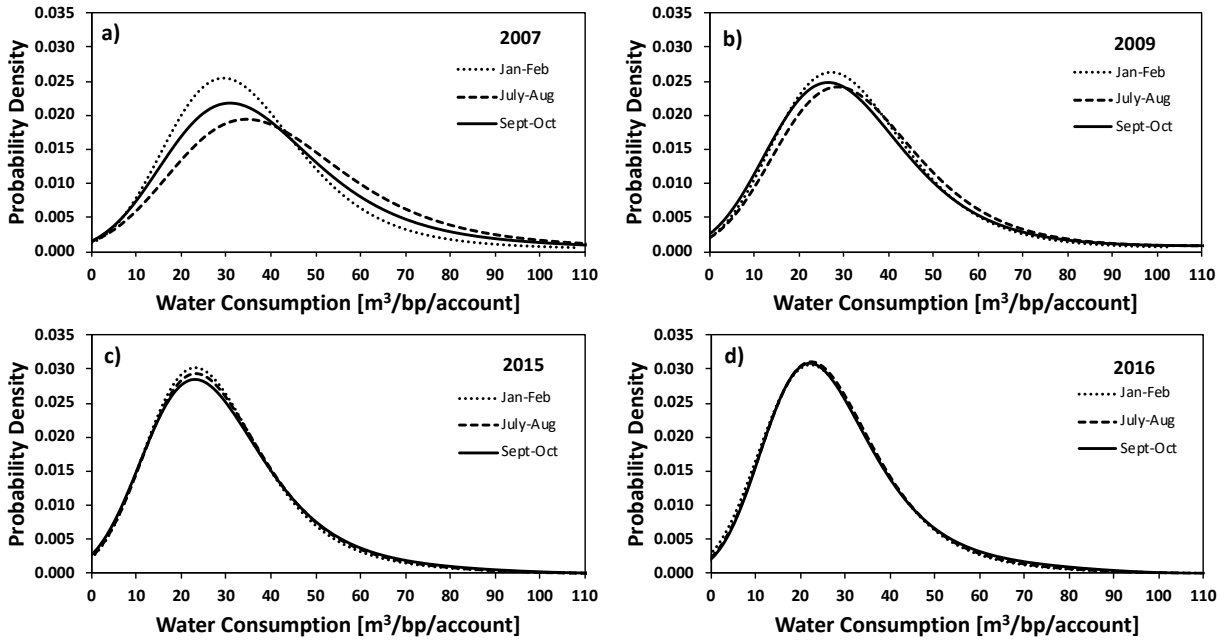

Figure 3.4: Select optimally parameterized PDFs showing seasonal influence on water consumption. Continuum response of the water consumption PDF to both real water price and weather using parametric PDFs from the January/February, July/August and September/October billing periods for 3.4.a) 2007, 3.4.b) 2009, 3.4.c) 2015, and 3.4.d) 2016.

The evolving co-dependence of consumer water consumption behaviour to changing weather conditions as well as the real water price within the City of Waterloo can be rationalized with two likely explanations. First, the economic disincentive for outdoor water use increases with an increase in the real water price, resulting in a diminished consumer response to weather

conditions. Second, the Region of Waterloo implemented their Water Efficiency Master Plan [2015-2025] (Region of Waterloo, 2014) influencing all billing periods in 2015 and onwards. This plan promotes water conservation through education, marketing, and controlling outdoor water usage through by law enforcement. Although the policy goals of the renewed master plan are consistent with the Water Efficiency Master Plan Update [2007-2015] (Region of Waterloo, 2006a), the recent addition includes a Residential Water Savings Assistance Program (RWSAP). Figures 3.4.c and 3.4.d suggest that the combination of economic disincentive and policy initiatives have negated any perceivable difference between the winter and summer water consumption histograms.

3.2.2. PRICE AND WEATHER AS AMBIENT PROCESSES

Previously, this chapter established that the ensemble residential accounts behave as a continuum process replicated by the water consumption PDF $p_{x,t}$. The next step is to quantify the correlation and infer causality in terms of how transient ambient processes influence the location $m_{x,i,t}$, scale $\sigma_{x,i,t}$, and shape parameters ($\alpha_{1,t}$, $\alpha_{2,t}$, $\alpha_{3,t}$, and $\alpha_{4,t}$) of the water consumption PDF $p_{x,t}$ as well as the corresponding mean statistic μ_x during each sampling period t. Ambient processes can be either macroscopic or microscopic in terms of their influence on consumers. Macroscopic ambient processes are experienced equally by all consumers within the utility, and include temperature, precipitation, real water price, education, and by-law enforcement. Intuitively, macroscopic processes should drive advection of the water consumption PDF through scaling of the median statistic. Additionally, they could also influence the scale and shape of the PDF provided the population of residential accounts experiences a heterogeneous response to changes in these utility-wide macroscopic processes. Microscopic processes are only experienced by a subset of the population and may include changes in household income and number of occupants. Microscopic processes may not have influence on a sufficient number of consumers to cause advection of the water consumption PDF. However, changes experienced by a subset of the population could influence dispersion through adjustments to the scale and shape of the PDF.

Price P_t is measured at each sampling period t and represents real water price as the depreciated variable unit cost of metered water. Prices are discounted using the annual consumer price index (CPI) inflation rate to a base year of 2004$. This analysis applies CPI under the assumption that it reflects increases in household income for all residential accounts, hence any

price increases above CPI represent real changes in water affordability relative to household income. The weather score W_t at each sampling period t is a function of rainfall R_t and temperature T_t measurements combined into a single process as:

$$W_t = T_t \times R_t \tag{3.15}$$

Where, T_t represents the average of the daily high temperature in degrees Celsius for all days within sampling period t (University of Waterloo Weather Station, 2017); and, R_t represents the number of days with less than 2mm of rainfall during sampling interval t (NASA, 2017; Environment Canada, 2017). The weather score is based on the hypothesis that temperature T_t and rainfall R_t are dependent variables that cannot be separated, and that changes in either cause a response in water consumption. Note that while the utility cannot control the weather score W_t, they are able to adjust the real water price P_t to ensure revenues generated from the variable unit cost of water promote financial sustainability. However, utilities adjust their water price once per year in advance of unknown seasonal weather variations within the target billing year. This minimizes the inter-dependence between the utility-controlled water price and seasonal variation in the weather score.

Figure 3.5 presents the discrete values for weather score and real water price for all billing periods between January/February 2007 and November/December 2016. The utility annually increases the real water price to boost their revenues, while the weather score changes periodically due to seasonal variability in temperature and precipitation. The troughs that appear along the weather score visualization represent the winter months, whereas the peaks represent summer months. Variability in the amplitude and width of the peaks are a consequence of seasonal weather variability that may include extreme weather events such as heavy rainfall in March/April and May/June or drought conditions in July/August and September/October billing periods. Appendix B.4 itemizes the ambient processes during each billing period.

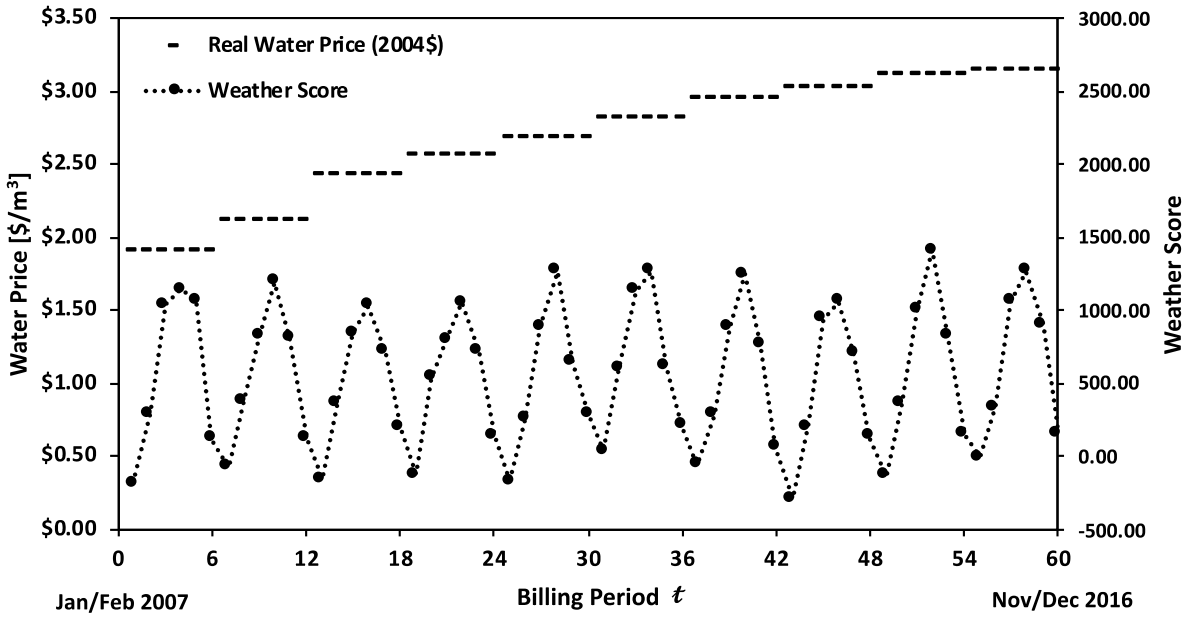

Figure 3.5: Ambient conditions for water consumption during entire analysis period. Time series representation of price P_t and weather score W_t variables for each bimonthly sampling period t.

3.2.3. ADVECTIVE-DISPERSIVE TRANSPORT MODEL PARAMETERIZATION

Previously, water consumption data from the City of Waterloo were used to establish the idea that the resulting PMF could be replicated with a parametric PDF, and evolution of this PDF represents the continuum response of the utility-wide residential accounts to real water price and weather conditions. The next step is to parametrize the coefficients within Equations 3.12 and 3.14 representing advective-dispersive transport of the PDF $\hat{p}_x$ and the mean statistic $\hat{\mu}_x$. Equations 3.10 and 3.11 show that these coefficients are partial derivatives of the location $m_{x,i,t}$, scale $\sigma_{x,i,t}$, and shape parameters ($\alpha_{1,t}$, $\alpha_{2,t}$, $\alpha_{3,t}$, and $\alpha_{4,t}$) with respect to real water price $\mathbb{x}_1 \sim P(t)$ and weather score $\mathbb{x}_2 \sim W(t)$. Advective-dispersive transport along with its partial-derivate coefficients are time-continuous process. This requires $t \to t$ resulting in a water consumption PDF $p_x(x, P, W)$ as well as its mean statistic $\mu_x(P, W)$ representing a transient one-dimensional process along the axis of water consumption x $[m^3/bp/acct]$. The probability $\mathcal{P}$ of any one residential consumer achieving a specified water consumption $\bar{x}$ can be found as $\mathcal{P} = \int_0^{\bar{x}} p_x(x, P, W)\, dx$ under predefined ambient conditions of $\mathbb{x}_1 \equiv P(t)$ and $\mathbb{x}_2 \equiv W(t)$ at time t.

Multi-variate curvilinear regression is used to estimate model parameters $b_{0\ldots7}$ within Equation 3.11 for the time-continuous statics $\widehat{U} \in \{\mu_x, m_x, \sigma_x, \alpha_{1\ldots4}\}$ for $\widehat{U}(b_{U,0\ldots7}, P, W)$. Estimation of model parameters $b_{0\ldots7}$ parameterizes the advective-dispersive transport model for the water consumption PDF $\hat{p}_x(x, P, W)$ and its resulting mean statistic $\hat{\mu}_{x|p_x}(P, W)$ given by Equation 3.12 and 3.13, respectively. Additionally, these model parameters also result in a description of the direct regression mean $\hat{\mu}_x(P, W)$ given by Equation 3.14. Table 3.2 summarizes results from the multi-variate curvilinear regression performed on the full suite of statistics listed in Appendix B.1 as well as the control function parameters $\alpha_{n_z,t}$ listed in Appendix B.2, with respect to the ambient conditions of real water price P_t and weather score W_t listed in Appendix B.4. Table 3.2 summarizes the curvilinear regression results of the general form of Equation 3.11 for the dependence of $\widehat{U} \in \{\mu_x, m_x, \sigma_x, \alpha_{1\ldots4}\}$ on $P(t)$ and $W(t)$, with model parameters $b_{U,0\ldots7}$ removed (set to "0.00") when p-values were greater than a 10% significance level.

Table 3.2: Summary of model parameterization for Equation 3.12.

U	$b_{U,0}$	$b_{U,1}$	$b_{U,2}$	$b_{U,3}$	$b_{U,4}$	$b_{U,5}$	$b_{U,6}$	$b_{U,7}$	Regression Analysis
$\hat{\mu}_x$	4.64×10^{1} ‡ $\begin{pmatrix} 2.36 \times 10^{1} \\ 8.71 \times 10^{-31} \end{pmatrix}$	$\begin{matrix} -5.45 \\ \begin{pmatrix} -7.54 \\ 4.38 \times 10^{-10} \end{pmatrix} \end{matrix}$	0.00	0.00	$\begin{matrix} 8.30 \times 10^{-6} \\ \begin{pmatrix} 3.32 \\ 1.61 \times 10^{-3} \end{pmatrix} \end{matrix}$	$\begin{matrix} -2.46 \times 10^{-6} \\ \begin{pmatrix} -2.71 \\ 8.86 \times 10^{-3} \end{pmatrix} \end{matrix}$	0.00	0.00	† $\begin{Bmatrix} 0.781 \\ 6.67 \times 10^{1} \\ 1.77 \times 10^{-18} \end{Bmatrix}$
$\hat{m}_x$	4.35×10^{1} $\begin{pmatrix} 2.51 \times 10^{1} \\ 1.95 \times 10^{-31} \end{pmatrix}$	$\begin{matrix} -5.41 \\ \begin{pmatrix} 8.43 \\ 2.04 \times 10^{-11} \end{pmatrix} \end{matrix}$	$\begin{matrix} -1.38 \times 10^{-2} \\ \begin{pmatrix} -1.83 \\ 7.32 \times 10^{-2} \end{pmatrix} \end{matrix}$	$\begin{matrix} 4.70 \times 10^{-3} \\ \begin{pmatrix} 1.70 \\ 9.44 \times 10^{-2} \end{pmatrix} \end{matrix}$	$\begin{matrix} 1.80 \times 10^{-5} \\ \begin{pmatrix} 2.70 \\ 9.30 \times 10^{-3} \end{pmatrix} \end{matrix}$	$\begin{matrix} -5.79 \times 10^{-6} \\ \begin{pmatrix} -2.40 \\ 1.97 \times 10^{-2} \end{pmatrix} \end{matrix}$	0.00	0.00	$\begin{Bmatrix} 0.816 \\ 4.80 \times 10^{1} \\ 1.16 \times 10^{-18} \end{Bmatrix}$
$\hat{\sigma}_x$	2.32×10^{1} $\begin{pmatrix} 1.32 \times 10^{1} \\ 8.13 \times 10^{-19} \end{pmatrix}$	$\begin{matrix} -2.05 \\ \begin{pmatrix} -3.17 \\ 2.50 \times 10^{-3} \end{pmatrix} \end{matrix}$	0.00	0.00	$\begin{matrix} 7.39 \times 10^{-6} \\ \begin{pmatrix} 3.30 \\ 1.68 \times 10^{-3} \end{pmatrix} \end{matrix}$	$\begin{matrix} -2.19 \times 10^{-6} \\ \begin{pmatrix} -2.70 \\ 9.13 \times 10^{-3} \end{pmatrix} \end{matrix}$	0.00	0.00	$\begin{Bmatrix} 0.571 \\ 2.48 \times 10^{1} \\ 2.37 \times 10^{-10} \end{Bmatrix}$
$\hat{\alpha}_1$	3.29×10^{-1} $\begin{pmatrix} 3.29 \times 10^{1} \\ 3.51 \times 10^{-39} \end{pmatrix}$	0.00	$\begin{matrix} 5.20 \times 10^{-5} \\ \begin{pmatrix} 3.90 \\ 2.51 \times 10^{-4} \end{pmatrix} \end{matrix}$	0.00	0.00	0.00	0.00	0.00	$\begin{Bmatrix} 0.208 \\ 1.52 \times 10^{1} \\ 2.51 \times 10^{-4} \end{Bmatrix}$
$\hat{\alpha}_2$	3.94×10^{1} $\begin{pmatrix} 4.75 \\ 1.50 \times 10^{-5} \end{pmatrix}$	$\begin{matrix} 1.47 \times 10^{-1} \\ \begin{pmatrix} 2.22 \\ 3.01 \times 10^{-2} \end{pmatrix} \end{matrix}$	0.00	$\begin{matrix} 1.27 \times 10^{-3} \\ \begin{pmatrix} 2.99 \\ 4.10 \times 10^{-3} \end{pmatrix} \end{matrix}$	0.00	$\begin{matrix} -8.27 \times 10^{-7} \\ \begin{pmatrix} -2.32 \\ 2.43 \times 10^{-2} \end{pmatrix} \end{matrix}$	$\begin{matrix} -3.24 \\ \begin{pmatrix} -2.50 \\ 1.53 \times 10^{-2} \end{pmatrix} \end{matrix}$	0.00	$\begin{Bmatrix} 0.359 \\ 7.69 \\ 5.37 \times 10^{-5} \end{Bmatrix}$
$\hat{\alpha}_3$	-0.658 $\begin{pmatrix} -2.55 \times 10^{1} \\ 7.80 \times 10^{-33} \end{pmatrix}$	0.00	0.00	$\begin{matrix} -2.12 \times 10^{-4} \\ \begin{pmatrix} -2.99 \\ 4.09 \times 10^{-3} \end{pmatrix} \end{matrix}$	0.00	0.00	0.00	$\begin{matrix} 5.99 \times 10^{-5} \\ \begin{pmatrix} 2.48 \\ 1.60 \times 10^{-2} \end{pmatrix} \end{matrix}$	$\begin{Bmatrix} 0.213 \\ 7.71 \\ 1.09 \times 10^{-3} \end{Bmatrix}$
$\hat{\alpha}_4$	7.89×10^{-2} $\begin{pmatrix} 1.14 \times 10^{1} \\ 1.70 \times 10^{-16} \end{pmatrix}$	0.00	0.00	$\begin{matrix} 1.10 \times 10^{-5} \\ \begin{pmatrix} 3.15 \\ 2.57 \times 10^{-3} \end{pmatrix} \end{matrix}$	0.00	0.00	0.00	0.00	$\begin{Bmatrix} 0.146 \\ 9.93 \\ 2.57 \times 10^{-3} \end{Bmatrix}$

Notes: Total degrees of freedom is 47 and residual degrees of freedom is 45.

‡ $\begin{pmatrix} \text{Regression Coefficient} \\ t \text{ statistic} \\ p-\text{value for } t \text{ statistic} \end{pmatrix}$ † $\begin{Bmatrix} R^2 \\ F \text{ statistic} \\ p-\text{value for } F \text{ statistic} \end{Bmatrix}$

The curvilinear regression results show that each of the mean $\hat{\mu}_x$, median $\hat{m}_x$, standard deviation $\hat{\sigma}_x$, and control function parameters $\hat{\alpha}_{1\ldots4}$ statistically correlate with the observed changes to P and W. The p-value on the F statistic indicates that there is less than 1% chance that any one relationship is coincidental, with the mean, median, and standard deviation showing stronger correlation than the control function parameters. Contributing parameters $b_{0\ldots7}$ vary between each statistic, which may indicate that the mean value as well as the location, scale, and shape of the distribution are controlled by different processes. Table 3.3 summarizes the active model parameters for each statistic as well as their derivative representation from the total derivative and Taylor series expansion given by Equations 3.8 and 3.9.

Table 3.3: Active model parameters for each statistic.

parameter		$\hat{\mu}_x$	$\hat{m}_x$	$\hat{\sigma}_x$	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$
$b_{U,0}$	$constant$	✓	✓	✓	✓	✓	✓	✓
$b_{U,1}$	$\frac{\partial U}{\partial P}$	✓	✓	✓	–	✓	–	–
$b_{U,2}$	$\frac{\partial U}{\partial W}$	–	✓	–	✓	–	–	–
$b_{U,3}$	$\frac{\partial^2 U}{\partial P \partial W}$	–	✓	–	–	✓	✓	✓
$b_{U,4}$	$\frac{\partial^2 U}{\partial W^2}$	✓	✓	✓	–	–	–	–
$b_{U,5}$	$\frac{\partial^3 U}{\partial P \partial W^2}$	✓	✓	✓	–	✓	–	–
$b_{U,6}$	$\frac{\partial^2 U}{\partial P^2}$	–	–	–	–	✓	–	–
$b_{U,7}$	$\frac{\partial^3 U}{\partial P^2 \partial W}$	–	–	–	–	–	✓	–

✓ *active model parameter*
– *inactive model parameter*

The R^2 value for the median statistic is larger than that for either the mean or the standard deviation, which indicates that the variance of the median is reproduced more accurately than for either the mean or standard deviation. Note that the standard deviation is dependent upon the of the median statistic (see Equation 3.5) and may inherit its estimation error. Contributing

coefficients $b_{0\ldots7}$ for the control function shape parameters ($\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\alpha}_3$, and $\hat{\alpha}_4$) vary and exhibit low R^2 values, which suggests that combinations of W and P only partially reflect observed variability. Three possible explanations exist for the inability of the curvilinear model to more accurately reproduce the observed variability in each statistic. First, unaccounted macroscopic ambient processes as well as microscopic household processes impact water consumption beyond that which can be explained by price and weather alone. These were itemized earlier as: passive water conservation, education, and by-law enforcement, as well as, household income and number of occupants. Second, imprecision of water consumption measurements recorded to within 1 m^3 severely restrict accuracy in the discrete median statistic $m_{x,i,t}$. This may influence the sensitivity of the model given that the discrete median $m_{x,i,t}$ is used to estimate the standard deviation $\sigma_{x,i,t}$ and each control function parameter $\alpha_{n_z,t}$, and ultimately influences the transformation between the measurement space x, the median-relative space y, and the standard score space z. This seems to have the greatest impact on estimating parameters for the control function model with their characteristic low R^2 values. Third, using time-averaged weather score data W_t that span bimonthly sampling periods may restrict the sensitivity of the curvilinear model from expressing severe and localized weather events. It is expected that shorter sampling intervals could provide greater resolution in the water consumption response to extreme seasonal weather conditions.

This advective-dispersive transport model (from Equation 3.12) representing the continuum response of the utility-wide residential water demand to the ambient process of real water price and weather score is now shown for the PDF solution $\hat{p}_x(x, P, W)$ as:

$$\begin{aligned}\hat{p}_x(x,P,W) = {} & \hat{m}_x(P,W) \\ & + \frac{1}{\hat{\sigma}_x(P,W)} \exp\left(-\int [\hat{\alpha}_1(P,W) + \hat{\alpha}_2(P,W)\, z + \hat{\alpha}_3(P,W)\, z^2 \right. \\ & \left. + \hat{\alpha}_4(P,W)\, z^3]\, dz\right)\end{aligned} \tag{3.16}$$

Similarly, the derived transport mean from Equation (3.13) $\hat{\mu}_{x|\hat{p}_x}(P,W)$ is expressed as:

$$\hat{\mu}_{x|\hat{p}_x}(P,W) = \hat{m}_x(P,W) + \hat{\sigma}_x(P,W)\int_{z_{min}}^{z_{max}} z\ \exp\left(-\int[\hat{\alpha}_1(P,W) + \hat{\alpha}_2(P,W)\,z + \hat{\alpha}_3(P,W)\,z^2 + \hat{\alpha}_4(P,W)\,z^3]\,dz\right)dz \tag{3.17}$$

where

$$\begin{aligned}
\hat{m}_x(P,W) &= b_{m,0} + b_{m,1}\,P(t) + b_{m,2}\,W(t) + b_{m,3}\,W(t)\,P(t) \\
&\quad + b_{m,4}\,W(t)^2 + b_{m,5}\,W(t)^2\,P(t) \\
\hat{\sigma}_x(P,W) &= b_{\sigma,0} + b_{\sigma,1}\,P(t) + b_{\sigma,4}\,W(t)^2 + b_{\sigma,5}\,W(t)^2\,P(t) \\
\hat{\alpha}_1(P,W) &= b_{\alpha_1,0} + b_{\alpha_1,2}\,W(t) \\
\hat{\alpha}_2(P,W) &= b_{\alpha_2,0} + b_{\alpha_2,1}\,P(t) + b_{\alpha_2,3}\,W(t)\,P(t) + b_{\alpha_2,5}\,W(t)^2\,P(t) \\
&\quad + b_{\alpha_2,6}\,P(t)^2 \\
\hat{\alpha}_3(P,W) &= b_{\alpha_3,0} + b_{\alpha_3,3}\,W(t)\,P(t) + b_{\alpha_3,7}\,P(t)^2\,W(t) \\
\hat{\alpha}_4(P,W) &= b_{\alpha_4,0} + b_{\alpha_4,3}\,W(t)\,P(t)
\end{aligned} \tag{3.18}$$

Finally, the direct regression mean from Equation (3.15) is now completed as:

$$\hat{\mu}_x(P,W) = b_{\mu,0} + b_{\mu,1}\,P(t) + b_{\mu,4}\,W(t)^2 + b_{\mu,5}\,W(t)^2\,P(t) \tag{3.19}$$

Substituting the partial derivatives from Table 3.3 representing the coefficients $b_{0\ldots7}$ into Equations 3.16, 3.17, 3.18, and 3.19 indicate that each statistic represents a partial differential equation. Moreover, the PDF $\hat{p}_x(x,P,W)$ itself is the solution to a partial differential equation representing the advective-dispersive transport of residential water demand in the three spatial dimensions x, P, W as well as time t.

3.3. DISCUSSION

Transport model regression results for the median $\hat{m}_x(P,W)$, standard deviation $\hat{\sigma}_x(P,W)$, and transport mean $\hat{\mu}_{x|\hat{p}_x}(P,W)$ are itemized in Appendix B.5 on Table B.5.1. The direct mean regression results are presented in Appendix B.5 on Table B.5.2. Visualization of models proceed

in the following order. Figure 3.6 depicts regression model ($\hat{m}_x$ and $\hat{\sigma}_x$) versus discrete ($m_{x,i,t}$ and $\sigma_{x,i,t}$) statics for the median and standard deviation as both time-series and in terms modeled versus actual values as a representation of model error. The issue of measurement accuracy of each water meter reading being restricted to 1 m^3 is clearly evident when viewing discrete measurements of the median $m_{x,i,t}$. Visualization of the data show a reasonably accurate fit which reinforces the correlation between measurements and the ambient variables quantified in Table 3.2, although the model clearly does not have sufficient sensitivity to exactly match the measured discrete statics. Notable outliers typically occur during the summer months and in the final year of the analysis. It is possible that the outliers are a reflection of sampling bias from collecting more data during bimonthly periods with lower weather scores such as the winter, spring, and fall seasons than in the summer. Also, the implementation of water consumption by-laws in the final two years of the analysis may have influenced how the model matches observational results in this period.

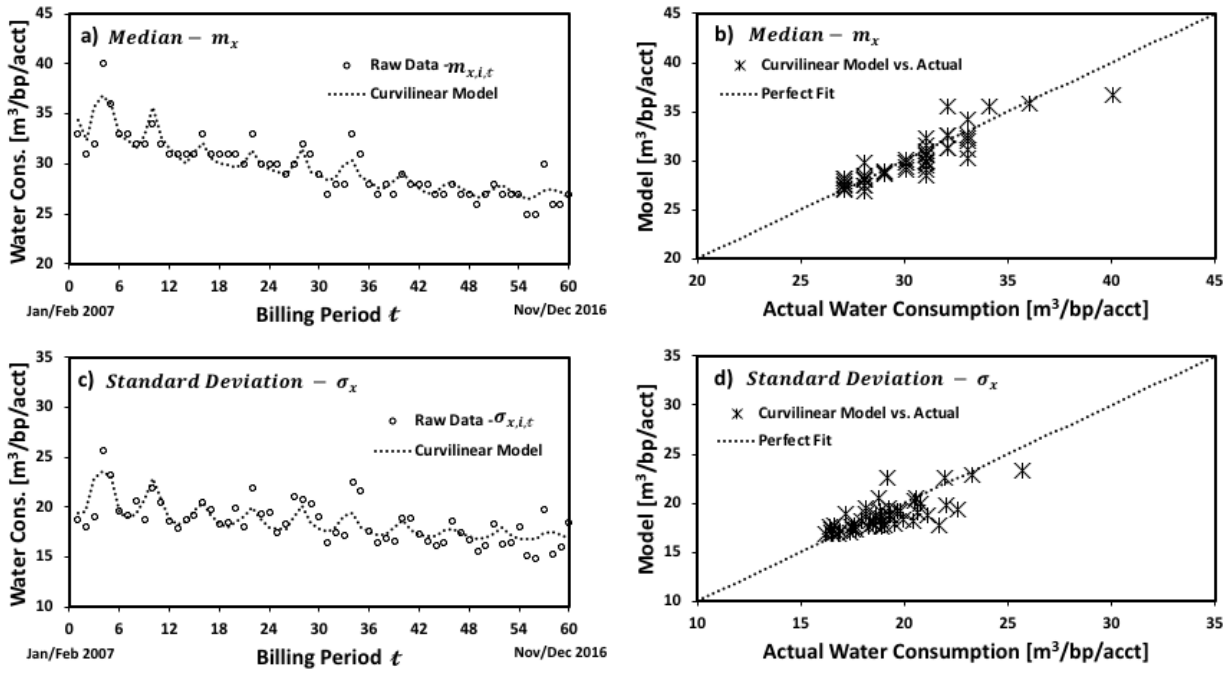

Figure 3.6: Curvilinear model results for the median and standard deviation. Visualization of the median $m_{x,i,t}$ and standard deviation $\sigma_{x,i,t}$ measurements and corresponding model results for $\hat{m}_x$ and $\hat{\sigma}_x$. Figures 3.6.a and 3.6.c present time-series visualizations of median and standard deviation model vs. measured values, while Figures 3.6.b and 3.6.d compare model to the realized median and standard deviation values, respectively.

Figure 3.7 depicts the late fall/winter November/December bimonthly histogram data from 2008 (3.7.a), 2010 (3.7.b), 2012 (3.7.c) and 2014 (3.7.d) normalized into PMFs (see Figure 3.1). Additionally, the parametric PDF arising from the optimal parametrization $p_{x,t}$ as well the advective-dispersive transport solution for $\hat{p}_x(x,P,W)$ given by Equation 3.16 are superimposed onto the PMFs. Finally, the transport mean $\hat{\mu}_{x|p_x}(P,W)$ given by Equation 3.17 as well as the arithmetic mean of the raw data $\mu_{x,i,t}$ are also presented. Figure 3.8 shows the same sequence of information except for the summer July/August bimonthly periods from 2007 (3.8.a), 2009 (3.8.b), 2011 (3.8.c) and 2013 (3.8.d). The advective-dispersive transport model for $\hat{p}_x(x,P,W)$ almost exactly reproduces the continuum response of residential water demand for the November/December periods which exhibits a low weather score, over the full range of real water price. The transport model becomes less accurate for the July/August summer months that correspond to a high weather score.

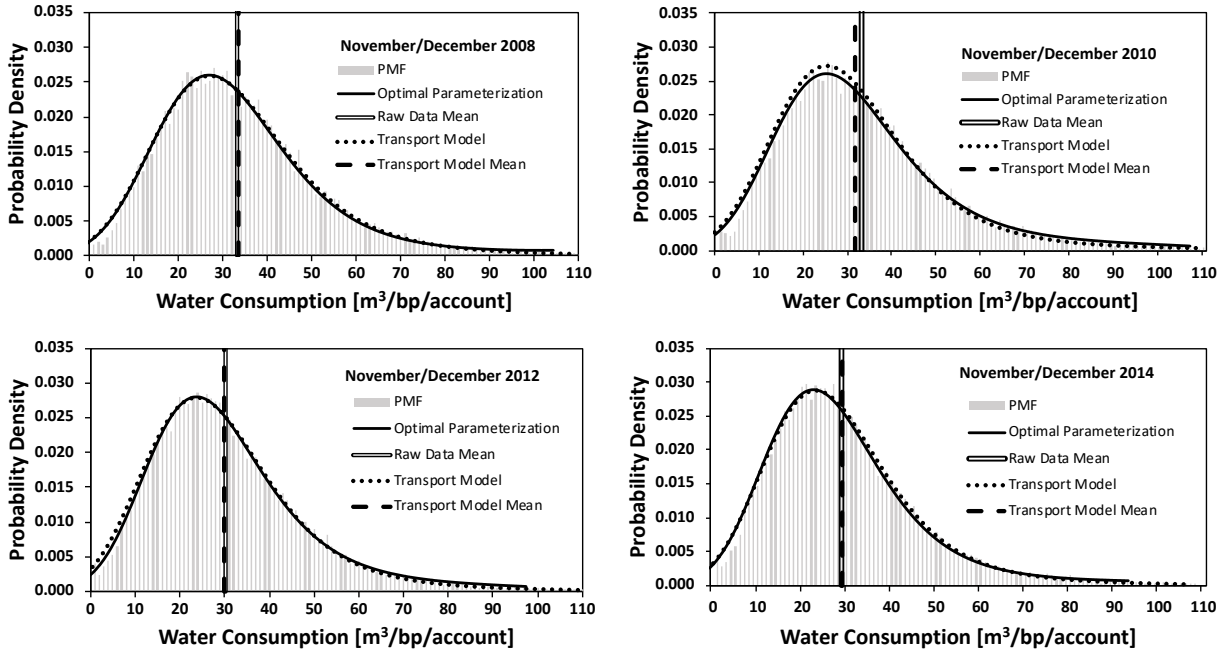

Figure 3.7: Transport model results for select November/December periods. Residential water consumption PMFs and corresponding PDFs for a sequence of November/December billing periods, with the optimal PDF from fitting the data, and PDF $\hat{p}_x$ obtained using the transport model. Also shown is the discrete mean $\mu_{x,i,t}$ as well as the estimated mean from the transport model $\hat{\mu}_{x|p_x}$.

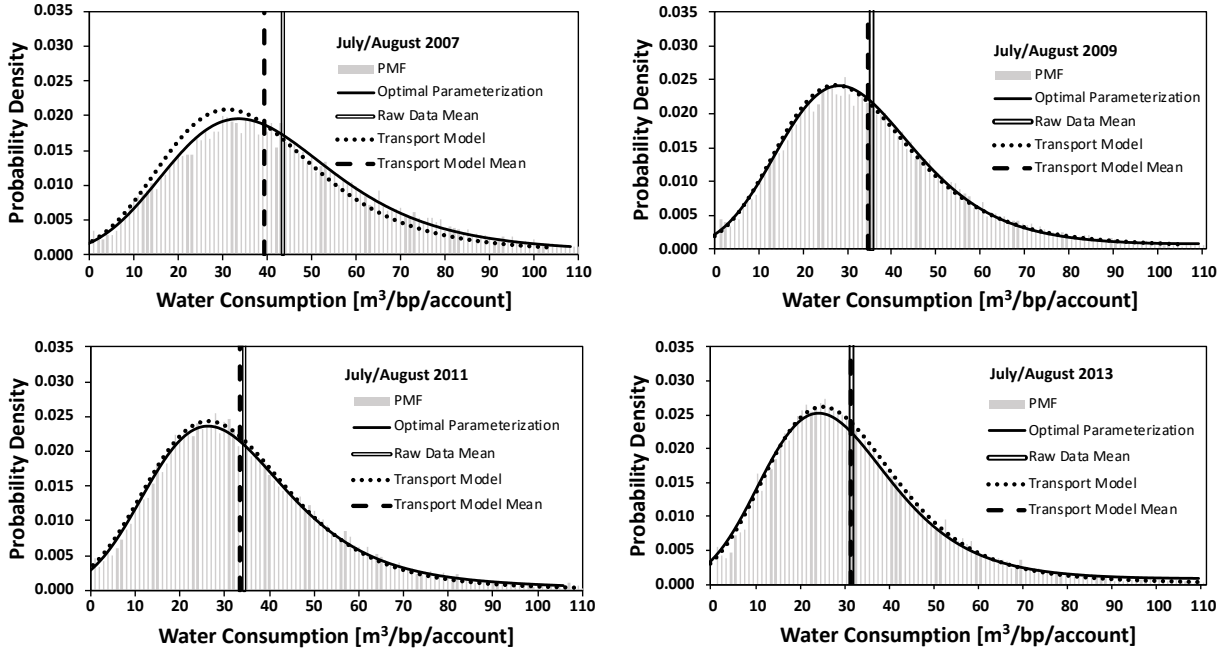

Figure 3.8: Transport model results for select July/August periods. Residential water consumption PMFs and corresponding PDFs for a sequence of July/August billing periods, with the optimal PDF from fitting the data, and the PDF $\hat{p}_x$ obtained using the transport model. Also shown is the discrete mean $\mu_{x,i,t}$ as well as the estimated mean from the transport model $\hat{\mu}_{x|p_x}$.

Figures 3.7 and 3.8 clearly indicate that as long as the continuum response of the entire system is adequately represented by $\hat{p}_x(x, P, W)$, then there is a unique representation for the transport mean $\hat{\mu}_{x|p_x}(P, W)$ such that it reproduces the arithmetic mean of the raw data. The direct regression mean $\hat{\mu}_x(P, W)$ given by Equation 3.20 does not include any information regarding the shape of continuum response (using the control function parameters) and is derived by observing how the arithmetic mean of the raw data $\mu_{x,i,t}$ directly responds to P_t and W_t. This independence to the control function provides an addition avenue for verification of the advective-dispersive transport process by comparing $\hat{\mu}_{x|p_x}$ with $\hat{\mu}_x$, as well as against $\mu_{x,i,t}$ for all sampling periods t. Values of $\hat{\mu}_{x|p_x}$ and $\hat{\mu}_x$ are itemized in Appendix B.5 on Tables B.5.1 and B.5.2, respectively. Additionally, they are visualized as a time series on Figure 3.9. Notice that the mean water consumption for July/August 2007 is somewhat underestimated by the transport model, perhaps

due to the issue of averaging short-duration extreme summer weather events over a two-month period to quantify the ambient process of weather score W_t. This is also observed during the 2010, 2012, and 2016 July/August bimonthly periods. However, both $\hat{\mu}_{x|p_x}$ and $\hat{\mu}_x$ exhibit nearly identical behaviour for all sampling periods. The transport mean $\hat{\mu}_{x|p_x}$is calculated by combining the location, scale, and shape of the continuum response as independent processes that all depend on $P(t)$ and $W(t)$. In contrast, the direct regression mean $\hat{\mu}_x$ does not differentiate between the location, scale and shape as it reproduces only the magnitude of the continuum response as a function of $P(t)$ and $W(t)$. Therefore, an analysis that only considers the direct regression model to assess a systems response naturally implies a loss of information.

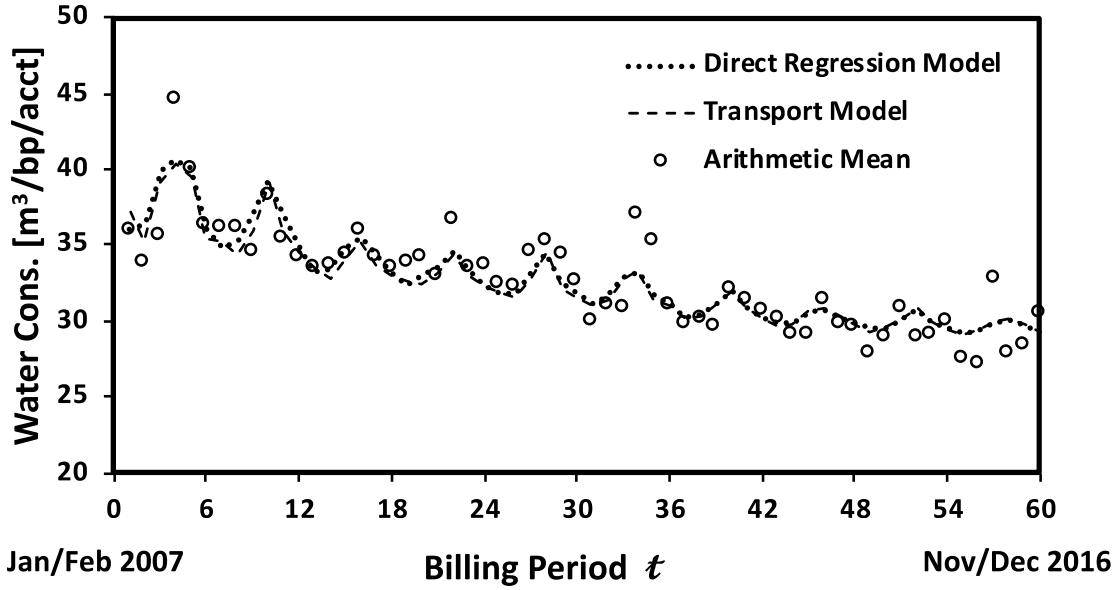

Figure 3.9: Transport model results for mean water consumption. The measured mean statistics $\mu_{x,i,t}$, direct regression model results $\hat{\mu}_x$, and the corresponding transport model results $\hat{\mu}_{x|p_x}$ for the entire analysis period.

PMFs from the summer billing periods of July/August 2015, May/June 2016, and July/August 2016 appear significantly different than previous years and suggest that omitted variables may be influencing water consumption. These are the only years where the measured mean water consumption in May/June (30.86 $m^3/bp/acct$ for 2015 and 32.88 $m^3/bp/acct$ for 2016) is higher than the water consumption in July/August (28.96 $m^3/bp/acct$ for 2015 and 27.95 $m^3/bp/acct$ for 2016) of the same year. Figure 3.10 shows that the transport model $\hat{p}_x(x, P, W)$

over-estimates the water consumption for the July/August period in both 2015 and 2016 despite that fact that the optimal parametrization $p_{x,t}$ is accurate. The hypothesis is that identifying and quantifying these potentially omitted variables could allow $\hat{p}_x(x, P, W) \rightarrow p_{x,t}$.

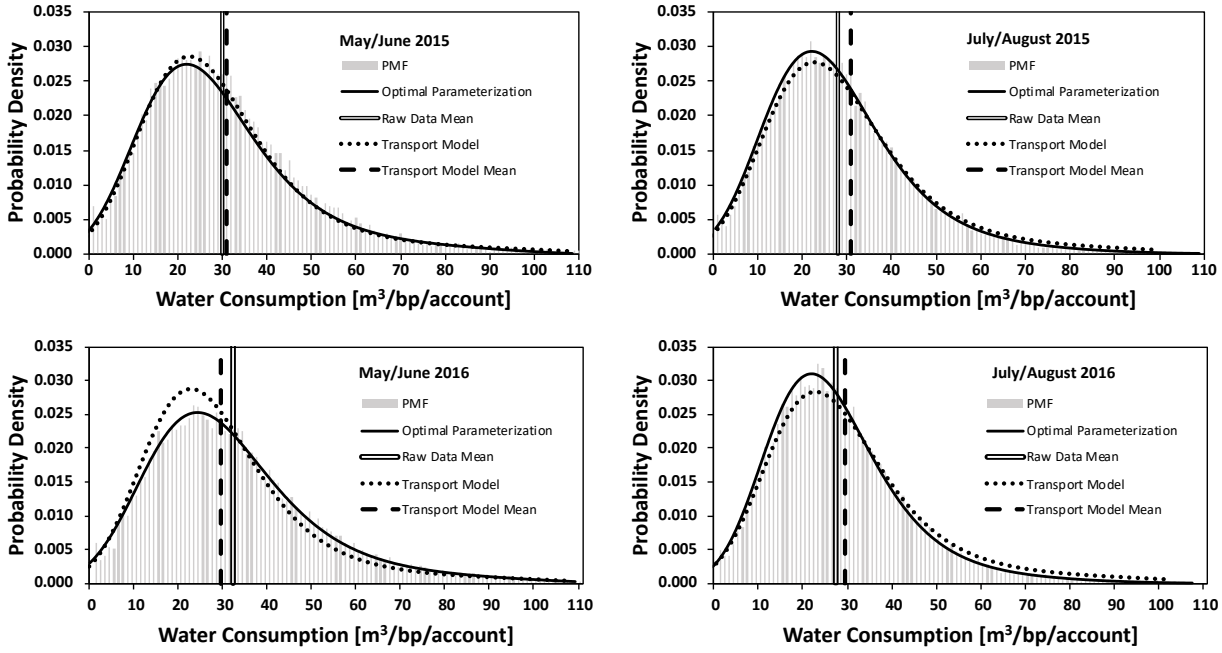

Figure 3.10: Transport model results for 2015-2016 May/June and July/August periods. Residential water consumption PMFs and corresponding PDFs for a sequence of May/June and July/August billing periods, with the optimal PDF from fitting the data, and the PDF $\hat{p}_x$ obtained using the transport model. Also shown is the discrete mean $\mu_{x,i,t}$ as well as the estimated mean from the transport model $\hat{\mu}_{x|p_x}$.

Starting in 2006, The Region of Waterloo implemented a *Water Efficiency Master Plan 2007-2015* which prioritized mass media advertising and a water conservation by-law that included outdoor water usage restrictions between May 31st and September 30th (Region of Waterloo, 2006b). Then, in 2015, The Region of Waterloo extended their water conservation practice with the *Water Efficiency Master Plan 2015-2025* including advertisement and enforcement. Additionally, they began to actively identify and target heavy water users through the RWSAP:

> "…who are known to have especially high household water use will be actively contacted and encouraged to participate in the program. This program will help address the challenge

shown frequently in market research that many residents are unaware that their consumption is markedly higher than the norm." (Water Efficiency Master Plan, 2015)

Beginning in May/June 2015, there appear to be tangible, yet unexplained, changes in residential water consumption habits that may indicate water conservation education and by-law enforcement act as ambient processes that significantly confound the influence of price and weather. For instance, the mean water consumption during the May/June billing period in 2015 was the highest since 2011 and actually increased from 2015 to 2016. Also, both the median and standard deviation were higher in May/June 2016 and were lower in July/August of the same year relative to what the advective-dispersive transport model could predict. This suggests the model under-anticipated consumption in May/June and over-anticipated consumption in July/August. Therefore, additional ambient processes arising from RWASP could include water conservation arising from improved education, awareness through mass media advertising, and active enforcement of water restriction by-laws may have led individual residential accounts to decrease their water usage during the July/August billing period. However, residents also appear to have substituted their water usage by increasing their water consumption in advance of the May 31st water restriction deadline prior to decreasing their consumption in the July/August period. Thereafter, even though they are conforming to policy, residential consumers return to historical water consumption patterns for the fall and winter months. In the context of this advection-dispersion analysis for the water consumption PDF and its mean statistic, it is seemingly impossible to measure policy as a continuum process in the same way as price P and weather score W. However, policy does appear to drive the continuum response of residential water consumption and this influence may be inferred through deviation from an otherwise accurate water consumption model.

3.4. CONCLUSIONS

The motivation for this study was to develop a methodology for water utilities to understand how changes to water price and weather patterns drive residential water consumption, and ultimately generate revenue to balance operational expenses in support of financial sustainability. This analysis involves histogram data representing residential water consumption within the City of Waterloo, Ontario, Canada over a 10-year period. The water consumption histograms are smooth, unimodal, asymmetric, shifted, with a heavy tail. They are transformed

into PMFs and shown to respond in a systemic manner to annual price increases as well as seasonal temperature and precipitation patterns. Following these observations, the premise of this analysis is that utility-wide residential water consumption of individual account holders exhibits a continuum response to changes in ambient variables such as price and weather.

To replicate the observed continuum response of the water consumption histogram, this chapter individually fits a parametric PDF to each of the 60 bimonthly histograms using a third-order exponential polynomial control function and reproduces the bimonthly data as a continuous function. The resulting PDFs as well their derived mean statistic represent the solution to an advection-dispersion transport equation, where: the median represents advection by locating the PDF and the standard deviation combined with the standard-score space PDF represents dispersion by virtue of giving the solution scale and shape. Optimally parameterizing the control function requires conservation of probability, which includes all water consumption measurements in the analysis to ensure the corresponding CDF is unity. Consequently, the advective-dispersive transport process is one-dimensional along the axis of water consumption $x\ [m^3/bp/acct]$. Therefore, the probability $\mathcal{P}$ that an account will achieve a specified water consumption $\bar{x}$ can be estimated numerically as $\mathcal{P} = \int_0^{\bar{x}} p_x(x, P, W)\, dx$ under predefined ambient conditions of real water price $P(t)$ and weather score $W(t)$ at time t.

The outcome of this analysis provides new possibilities for interpreting how the location, scale, and shape of a distribution of measurements respond to changes in ambient conditions. This analysis demonstrates that it is reasonable to disaggregate the data into "advection-dispersion" like components to characterize how a distribution will evolve through time. Furthermore, this approach guarantees a unique solution for the mean statistic and may provide more precision when forecasting how the solution will evolve for future known or anticipated ambient conditions. This could have positive implications for water utilities attempting to forecast their revenues under a strict water price increase schedules in the face of ever-increasing expenses. Furthermore, this approach could provide utilities with the ability to quantify the influence of policy implementation and enforcement such as summer water use restrictions. Additionally, this advective-dispersive transport framework for consumer behaviour could analogously apply to other industries such as electrical utilities and transportation services, as well as, social and health science applications with histogram data that exhibit non-Gaussian tendencies.

4. Conservation of Probability and Parametric PDEs

The first two main chapters of this thesis outline a methodology for transforming discrete data histograms into continuous probability density functions (PDFs) and evaluating how these functions evolve through time. The outcome of Chapter 3 is a governing partial differential equation (PDE) that describes how the water consumption PDF will evolve with respect to water price and weather conditions. The water consumption data provide a useful example for presenting any distribution of discrete measurements as a probabilistic process using the following concepts: conservation of probability, spatial-continuity, and temporal-continuity. Anecdotal application to water consumption data motivates the investigation of an overarching theory that describes how information evolves within both physical and abstract systems. This chapter attempts to articulate a general understanding from this specific application.

The water utility application tracks the evolution of residential water consumption PDFs in bimonthly intervals over a 10-year period. Price and weather are identified as ambient conditions that likely influence the water consumption PDF. Curvilinear regression supports a statistically significant correlation between the median, standard deviation, and control function parameters that recombine to describe the evolution of the water consumption PDF. The curvilinear regression parameters represent partial derivatives that describe the relationship between the median, standard deviation, and control function with the ambient conditions. Therefore, the outcome of the water consumption application is a PDE for each statistic with respect to the ambient conditions of price and weather as they evolve through time. By combining the PDEs for each statistic in Equation 3.1 the PDE resulting from the transport model solves to reproduce a PDF. This is consistent with the understanding that a normal distribution represents the solution to a second-order homogeneous PDE. Solving the transport model for alternative forms of PDEs may provide insight into systems that are not governed by the normal distribution.

The goal of this investigation is to build an understanding of this overarching framework using probabilistic systems that are both simpler than the water consumption application and also well understood. Therefore, this framework we examine the concepts of probability, spatial-continuity, and temporal-continuity in the context of mass transport through molecular diffusion. Using Einstein's second order homogeneous PDE for molecular diffusion, this chapter provides a probabilistic derivation of molecular self-diffusion in two dimensions, which is consistent with the

Fourier solution. Fourier's heat conduction equation is a PDE and its solution is constrained to be a scaled normal distribution. This solution is a versatile analytical relationship that analogously applies to disparate and seemingly unconnected disciplines. Narasimhan (1999) provides an insightful overview of the history, influence, and applications of this solution within contemporary science. The influence of Fourier's solution is best described through advancement in the fields of electricity, flow in porous media, and molecular diffusion. Although these fields are seemingly disparate, historical modeling of experimental observations using the Fourier solution suggests that it mathematically describes specific processes within each field. Moreover, it may represent an overarching conservation law that governs how information flows within physical systems. A natural extension of this derivation produces Fick's Law and reveals that the diffusion coefficient is inversely proportional to the component density of the fluid medium.

This chapter provides compelling evidence that conservation of probability applies to disparate fields, which suggests conservation of information is a unifying concept for modeling systemic response. The generality of this approach to govern both societal and physical processes allows this analysis to conclude that second-order PDEs reflect interactions between three well-defined system components: 1) a source/sink term, 2) the measured property, and 3) a conduit that connects the source/sink to the measured property. This realization indicates that parameters within the governing PDE meaningfully describe and quantify the properties of the conduit. For the water consumption application, the conduit is the household-specific qualities that compel water consumption, the source term is the necessity to consume water to maintain standard of living conditions, and the measurement is the volume of water consumed. Future work could exploit parameterization of the PDE to forecast pairs of source/sink terms and the resulting solution of a measurement PDF.

4.1. THEORY

Conservation of probability, spatial-continuity, and temporal-continuity provide the foundation for interpreting governing PDEs for various physical and abstract processes. Conservation of probability refers to the sum of probability not exceeding unity for all potential outcomes. Spatial-continuity refers to a continuous function that describes the probability for any discrete interval within the measurement space. Temporal-continuity refers to the spatially-continuous PDF also being continuous with respect to time. This chapter largely focuses on

interpreting experimental results as the solution to a set of governing PDEs that describe how the distribution of experimental results will evolve through time as a function of ambient processes. In this context, experimental observations can be categorized as either intensive or extensive with respect to both space and time. According to the International Union of Pure and Applied Chemistry (IUPAC), intensive properties in space describe the governing system but do not depend on the amount of substance that they describe, while extensive properties in space are additive and correspond directly to the amount of substance they describe. This chapter discusses how the general concepts of intensive and extensive measurements relate to the water consumption application.

Figure 4.1.a presents water consumption measurement data form July/August 2007 in extensive $\mathcal{E}$, intensive $\mathcal{I}$, parametric PDF, and parametric CDF representations. Figure 4.1.a shows a histogram where the vertical axis represents the number of accounts as an extensive measurement $\mathcal{E}\ [frequency]$ and the horizontal axis represents water consumption $x\ [m^3/bp]$, where 'bp' is bimonthly period. The histogram bins in Figure 4.1.a represent the number of accounts that consume water within the time- and space-interval of $\mathcal{E} = \int_{t_1}^{t_2}\int_{x_1}^{x_2}\int \mathcal{I}_{x,t,\mathcal{N}}\, d\mathcal{N}\, dx\, dt$, where $\mathcal{N}$ is the number of measurements, dt is a time-interval for measurement, and dx is the horizontal bin size. The number of extensive measurements is transformed into an intensive representation by $\frac{d\mathcal{E}}{d\mathcal{N}} = \int_{t_1}^{t_2}\int_{x_1}^{x_2} \mathcal{I}_{x,t,\mathcal{N}}\, dx\, dt$, which also converts the vertical axis of the histogram from account frequency to probability density. This is expressed in the transition from Figure 4.1.a to Figure 4.1.b.

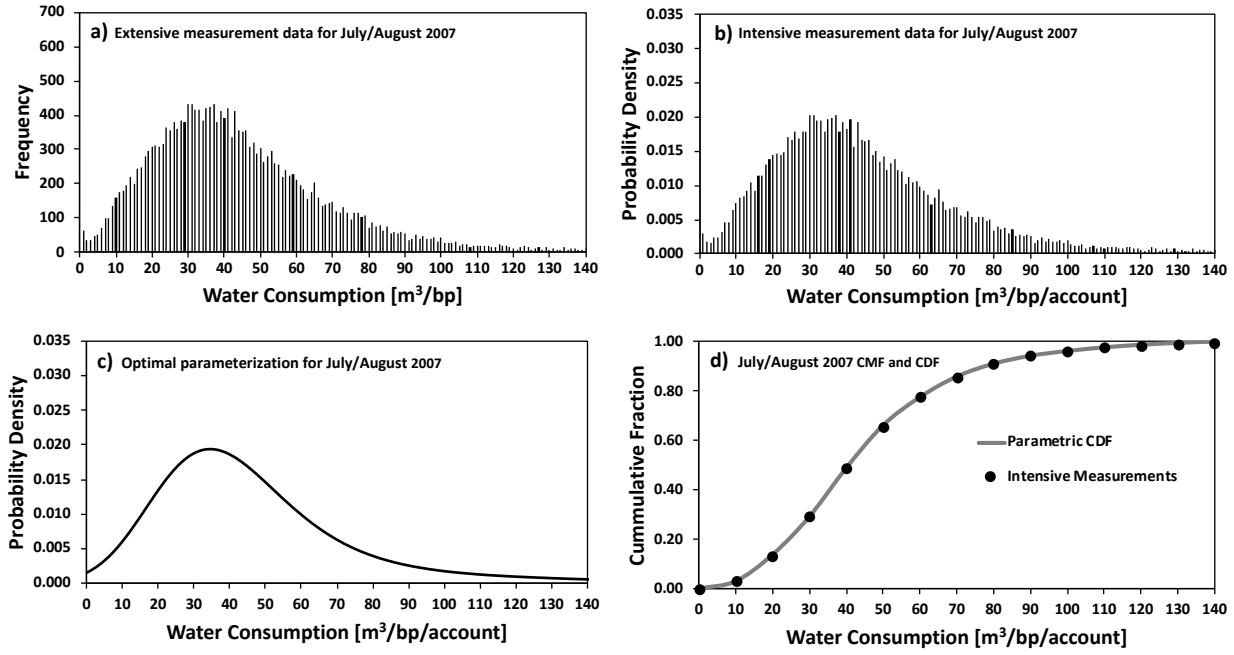

Figure 4.1: Water consumption data and parametric PDF for July/August 2007.

The relationship between intensive and extensive data implies conservation of probability, because the intensive distribution accounts for all of the extensive measurements. Summing the probability bins of intensive data from $0 \rightarrow \infty$ in Figure 4.1.b naturally captures all possible measurements and results in a cumulative probability of unity. Figure 4.1.c shows the derived intensive PDF representation of the original extensive histogram data. Integrating the PDF over the range of x over some sampling interval t results in a CDF shown on Figure 4.1.d that conserves probability. Evaluating the position-weighted frequency of the intensive and extensive histogram data provide an estimate of the mean occurrence $\mu_x = \int x\mathcal{I}_{x,t,\mathcal{N}}\, dx$ and $\mu_x = \int x\mathcal{E}_{x,t}\, dx$. Further integrating the mean occurrence by the number of measurements estimates the total measure of the system as: $\mathcal{M}_x = \int \mu_x\, d\mathcal{N}$. For the water consumption histogram, the total measure of the system is the utility-wide residential water consumption, which is important for developing sustainable utility management practices. This result is the culmination of conservation of probability, spatial-continuity, and temporal-continuity, which allows this theory to deconstruct a highly-complex water consumption problem into manageable concepts.

Equation 4.1 generalizes the above principles and presents conservation laws in the context of intensive and extensive properties. Notably, the frequency representation of data from Figure

4.1 can be used to infer the properties of a PDF through conservation of probability, $\int \mathcal{I}_{x,t,\mathcal{N}}\, dx = 1 \iff \mathcal{I}_{x,t,\mathcal{N}} = 0$. In this context, the intensive property $\mathcal{I}_{x,t,\mathcal{N}}$ is simultaneously an expression of space x, time t, and frequency, normalized by the total number of measurements $\mathcal{N}$ as:

$$Conservation\ of\ Probability, \qquad \int \mathcal{I}_{x,t,\mathcal{N}}\, dx = 1 \iff \mathcal{I}_{x,t,\mathcal{N}} = 0$$

$$Spatial - Continuity, \qquad \mathcal{E}_t = \iint \mathcal{I}_{x,t,\mathcal{N}}\, dx d\mathcal{N} \iff \mathcal{I}_{x,t,\mathcal{N}} = \frac{d^2\mathcal{E}_t}{dx d\mathcal{N}} \tag{4.1}$$

$$Temporal - Continuity, \qquad \mathcal{E} = \iiint \mathcal{I}_{x,t,\mathcal{N}}\, dx d\mathcal{N} dt \iff \mathcal{I}_{x,t,\mathcal{N}} = \int \frac{d}{dt}\left[\mathcal{I}_{x,t,\mathcal{N}}\right] dt$$

Equation 4.1 provides a theoretical foundation for investigating the relationship between probabilistic PDEs and their associated PDF solutions. The relationships in Equation 4.1 specify subscript notation for both intensive $\mathcal{I}$ and extensive $\mathcal{E}$ measurements through differentiation and integration. Each subscript represents a derivative of the extensive measurements with respect to space x, time t, and total measurements $\mathcal{N}$. This notation clearly communicates the relationship between extensive measurements, the resulting intensive PDF, and the parametric representation in Figure 4.1. The analysis in chapter 3 applies conservation of probability by comparing the CMF to a CDF (Figure 4.1.d) to fit a parametric PDF (Figure 4.1.c) to the intensive measurements (Figure 4.1.b). Thus, the analysis uses these foundational concepts to evaluate how the water consumption distribution evolves with time. The concept of time-continuity suggests that the measurement PDF can continuously transition from one steady-state condition to another. In this context, the results in Figure 4.1 provide a time-still snapshot of water consumption during the July/August 2007 bimonthly period.

Conservation of probability ensures that all measurements are accounted for and allows the PDF of intensive properties to rescale itself accordingly as the system changes. This is especially important for evaluating the transition from one steady-state condition to another, to progressively account for the dynamic nature of a system. Spatial-continuity is important for interpreting measurements in the context of conservation of probability. Normalizing the frequency $\mathcal{E}_{x,t}$ by the total number of measurements $\mathcal{N}$ for defined space and time intervals naturally transforms empirical data to represent a percentage of total measurement – which provides a suitable definition for an intensive property $\mathcal{I}_{x,t,\mathcal{N}}$. Summing the intensive properties over the entire

measurement space, for example the water consumption data in Figure 4.1, will always sum to unity through $\int \mathcal{I}_{x,t,\mathcal{N}}\, dx = 1$. However, the intensive distribution may change with respect to time, implying $\frac{d\mathcal{I}_{x,t,\mathcal{N}}}{dt}$ exists and is non-zero.

The example data provided in Figure 4.1 do not have the characteristics of a normal distribution. Chapter 3 investigates the evolution of the water consumption PDF and develops a PDE that accurately reproduces this PDF over a 10-year period. This chapter contends that characterizing PDEs that can reproduce the distribution of observation data is a form of data compression. This chapter provides a general methodology interpreting a multitude of data measurements and storing them within PDE parameterization that solves for the measurement distribution. The context for data compression is the storage of many data point using less PDE parameters without losing data fidelity. Although the water consumption application provides an adequate example of the relationship between PDEs and PDFs, there is merit to showing this relationship for the normal distribution. The following sections continue from the basis second-order homogeneous PDEs which solve to produce a normal distribution. This chapter considers the example process of molecular self-diffusion and its second-order homogeneous PDE.

4.1.1. NORMALLY DISTRIBUTED MEASUREMENTS

It is well-understood that the normal distribution represents the solution to a second-order homogeneous PDE. Here, the analysis reimagines a PDE that uses the notation for intensive measurements, consistent with Figure 4.1.b, that combine to produce the shape of a normal distribution. Here, the intensive measurements $\mathcal{I}_{x,t}$ represent a distribution in both space and time as:

$$\frac{d}{dt}\left[\mathcal{I}_{x,t,\mathcal{N}}\right] = D_{x,t,\mathcal{N}}\frac{d^2}{dx^2}\left[\mathcal{I}_{x,t,\mathcal{N}}\right] \tag{4.2}$$

Where, $D_{x,t,\mathcal{N}}$ is the space-time scaling coefficient, which is necessary to enforce dimensional consistency. Notably, Einstein (1905) uses this same form to characterize molecular diffusion in one dimension through a Taylor series approximation. As a second-order homogeneous PDE, the solution to Equation 4.2 is naturally defined as:

$$\mathcal{I}_{x,t,\mathcal{N}} = \frac{1}{\sqrt{s^2_{x,t,\mathcal{N}}}}\bar{p}_z \implies \frac{dp^*_{x,t}}{d\mathcal{N}} = \mathcal{I}_{x,t,\mathcal{N}}, \qquad \therefore \frac{dp^*_{x,t}}{d\mathcal{N}} = \frac{1}{\sqrt{s^2_{x,t,\mathcal{N}}}}\bar{p}_z \tag{4.3}$$

Where, $\frac{dp^*_{x,t}}{d\mathcal{N}}$ represents the intensive distribution of measurements, the intensive property $\mathcal{I}_{x,t,\mathcal{N}}$ is naturally a PDF due to its dependence on $p^*_{x,t}$ which in itself is a scaled PDF derived from the space- and time-invariant standard-score normal distribution $\bar{p}_z$, $s^2_{x,t,\mathcal{N}}$ is the variance of the intensive property. Note the bar notation '–' expresses the standard-score PDF as a normal distribution $\bar{p}_z$. The solution to the PDE in Equation 4.2 is scaled by the inverse square-root of the variance $\frac{1}{\sqrt{s^2_{x,t,\mathcal{N}}}}$ and the * superscript for the PDF $p^*_{x,t}$ represents a zero-centered distribution about the median $m_{x,t,\mathcal{N}}$ in the measurement space x. Using $\mathcal{E}_{x,t} = \int \mathcal{I}_{x,t,\mathcal{N}}\, d\mathcal{N}$, this analysis can express the intensive property $\frac{dp^*_{x,t}}{d\mathcal{N}}$ as a distribution of extensive measurements by integrating over the total number of measurements $\mathcal{N}$. The resulting dimensionality of the intensive property $\frac{dp^*_{x,t}}{d\mathcal{N}} = \mathcal{I}_{x,t,\mathcal{N}}$ is consistent with probability density, where it expresses the number of measurements within an infinitesimally small bin in the measurement interval dx, over time-interval dt, with respect to the total number of measurements $\mathcal{N}$.

Interpreting the measurements that describe physical systems is important in the context of developing and parameterizing PDEs. For instance, this investigation focuses on second-order homogeneous PDEs for diffusion as expressed in Einstein, which were experimentally confirmed by Perrin (1913) and also Fick's Law, which is later proposed to be an extension of Einstein. Table 4.1 introduces the second-order homogeneous PDEs for Einsteinian diffusion, and Fick's Law, which produce solutions that reflect a scaled normal distribution through $\frac{dp^*_{x,t}}{d\mathcal{N}} = \frac{1}{\sqrt{s^2_{x,t,\mathcal{N}}}}\bar{p}_z$.

Table 4.1: Second-order Homogeneous PDEs for various physical processes.

Application	Time-derivative Intensive PDF	Spatial-derivative Intensive PDF	Scaling Coefficient
Einsteinian Diffusion	$\frac{d}{dt}\left[\frac{dp^*_{\ell,t}}{dn}\right]$	$\frac{d^2}{d\ell^2}\left[\frac{dp^*_{\ell,t}}{dn}\right]$	$D_{\ell,t,n} = \frac{\partial^2 \ell}{\partial t^2} dt$
Fick's Law	$\frac{d}{dt}\left[\frac{dn}{dV}\right]$	$\frac{d^2}{dr^2}\left[\frac{dn}{dV}\right]$	$D_{n,t,V} = \frac{\partial^2 r}{\partial t^2} dt$

Where, $p^*_{\ell,t}$ represents the distribution of molecular displacement $[length]$, n represents molecules $[molecules]$, ℓ represents displacement magnitude $[length]$, $D_{\ell,t,n}$ is the Einsteinian diffusion coefficient $[length^2/time]$, V represents volume $[length^3]$, r represents radial displacement $[length]$, $D_{n,t,V}$ is Fick's Law diffusion coefficient $[length^2/time]$.

From Table 4.1, notice that the "time-derivative intensive PDF" are located on the left-hand side of Equation 4.2 as $\frac{d}{dt}\left[\mathcal{I}_{x,t,\mathcal{N}}\right]$. The "spatial-derivative intensive PDF" are located on its right-hand side of this equation as $\frac{d^2}{dx^2}\left[\mathcal{I}_{x,t,\mathcal{N}}\right]$. A time-dependent second-order homogeneous PDE implies the transient evolution of a spatial distribution. The processes in Table 4.1 require extensive measurements of displacement and moles to describe the PDEs for Einstein and Fick, respectively. However, the PDEs require intensive expressions of these measurements, which divide by additional extensive measurements of moles and volume. Dividing one extensive measurement by another produces a derived intensive property, known as a composite property, that applies to the PDEs in Table 4.1 as: $\mathcal{I}_{x,t,\mathcal{N}} = \frac{dp^*_{\ell,t}}{dn}$ and $\mathcal{I}_{x,t,\mathcal{N}} = \frac{dn}{dV}$.

4.1.2. SCALING THE BROWNIAN MOTION DIFFUSION COEFFICIENT

The intent of this section is to investigate the process of diffusion as a solution to Einstein's probabilistic PDE through an interpretation of intensive and extensive properties. Perrin's (1913) measurements of molecular diffusion provide motivation for this investigation to rescale Einstein's PDE to reflect two-dimensional diffusion. Using the general form PDE from Equation 4.2, this investigation will show that transforming the molecular displacement PDF from linear-Cartesian to linear-polar coordinates requires rescaling of the diffusion coefficient to conserve probability.

In this spirit, this investigation proceeds to evaluate diffusion processes that conserve probability in two-dimensions. The outcome of this section is a probabilistic solution to two-dimensional diffusion that is consistent with the solution to the Fourier equation describing heat transfer in solids. Furthermore, this chapter shows that this interpretation may lead to a probabilistic derivation of Fick's Law.

4.1.2.1. One-Dimensional Diffusion

Einstein shows that a scaled normal distribution is the solution to the physical process of molecular diffusion that results from a second-order homogeneous PDE. This solution provides a relationship between the diffusion coefficient and microscopic movement expressed through a nonzero, time-dependent variance of molecular displacement. Expressing the diffusion PDE through the general form in Equation 4.2 requires that the distribution of molecular displacement $\frac{dp^*_{\ell,t}}{dn}$ is the intensive property as $\mathcal{I}_{x,t,\mathcal{N}}$; total number of moles n in a spatial and temporal interval is in the denominator of the intensive property as $\mathcal{N}$; and, the diffusion coefficient $D_{\ell,t,n}$ is the space-time scaling coefficient as $D_{x,t,\mathcal{N}}$.

$$\frac{d}{dt}\left[\frac{dp^*_{\ell,t}}{dn}\right] = D_{\ell,t,n}\frac{d^2}{d\ell^2}\left[\frac{dp^*_{\ell,t}}{dn}\right], \qquad \frac{dp^*_{\ell,t}}{dn} = \frac{1}{\sqrt{s^2_{\ell,t,n}}}\bar{p}_z, \qquad s^2_{\ell,t,n} = 2D_{\ell,t,n}t \tag{4.4}$$

Where, Equation 4.4 applies the Einstein-Smoluchowski equation to relate the diffusion coefficient to variance of molecular velocity; and, $\sqrt{s^2_{\ell,t,n}}$ represents the standard displacement or square root of variance of linear displacement in one dimension. The zero-centered PDF $\frac{dp^*_{\ell,t}}{dn}$ characterizes the frequency of displacement per molecule in a one-dimensional space. This interpretation represents one-dimensional diffusion at the microscopic level of magnitude ℓ in a positive or negative direction with respect to time t. Perrin (1913) confirmed that the time-dependent displacement of molecular movement is characterized by a normal distribution and supports Einstein's solution to the diffusion PDE. When divided by the total number of measurements, Perrin's experimental results are consistent with frequency of displacement per molecule per time $\frac{d}{dt}\left[\frac{dp^*_{\ell,t}}{dn}\right]$ which is an expression of molecular velocity. However, Perrin measures molecular displacement of suspended solids in two spatial dimensions (i.e. a petri dish), which suggests there should be an analogous

two-dimensional solution to Einstein's probabilistic PDE. Using this basis, the following sections use conservation of probability to rescale Einstein's displacement PDF to reflect a two-dimensional solution in polar coordinates.

An example may help to visualize the link between Perrin's experiment and the probabilistic solution to Einstein's PDE. Ultimately, Perrin measures the magnitude and angle of displacement resulting from microscopic displacement of macro-particles that is consistent with Brownian motion while suspended in a liquid. Interpreting the displacements as either positive or negative while ignoring the angle of displacement produces a binomial normal distribution. However, the physical process of diffusion requires both a magnitude and direction to be consistent with the realized experimental results. Therefore, this section contends that combining a normal distribution of displacement ℓ and a uniform distribution of direction θ can recreate an example distribution that is consistent with Perrin's results. Randomly selecting a step-size and direction from these distributions using a consistent interval of time produces a two-dimensional random walk. Figure 4.2 presents an example illustration of molecular velocity, which produces results that are consistent with those of Perrin.

Representing a two-dimensional random walk using sequential steps in time (Figure 4.2.a) is analogous to the random walk envisioned by Einstein to develop his probabilistic PDE. However, Einstein effectively assumes radial symmetry for his derivation without explicitly stating this assumption. Grouping the velocity measurements by quadrant $(++, -+, +-, --)$, this analysis can interpret the example displacements from Figure 4.2 in one-dimension. This exercise groups steps taken within quadrants $(++, -+)$ and considers them to represent positive displacements – these data points are positioned between $0 \leq \theta \leq \pi$ on Figure 4.2.a. Furthermore, this exercise groups steps taken within quadrants $(+-, --)$ and considers them to be negative displacements – these data points are positioned between $\pi \leq \theta \leq 2\pi$ on Figure 4.2.a. Combining the positive and negative displacement values results in a binomial distribution that is equivalent to the one-dimensional distribution envisioned by Einstein. Given that Einstein's one-dimensional PDF is supported by two-dimensional measurements, this suggests that there exists a two-dimensional probabilistic solution for molecular diffusion that conserves probability.

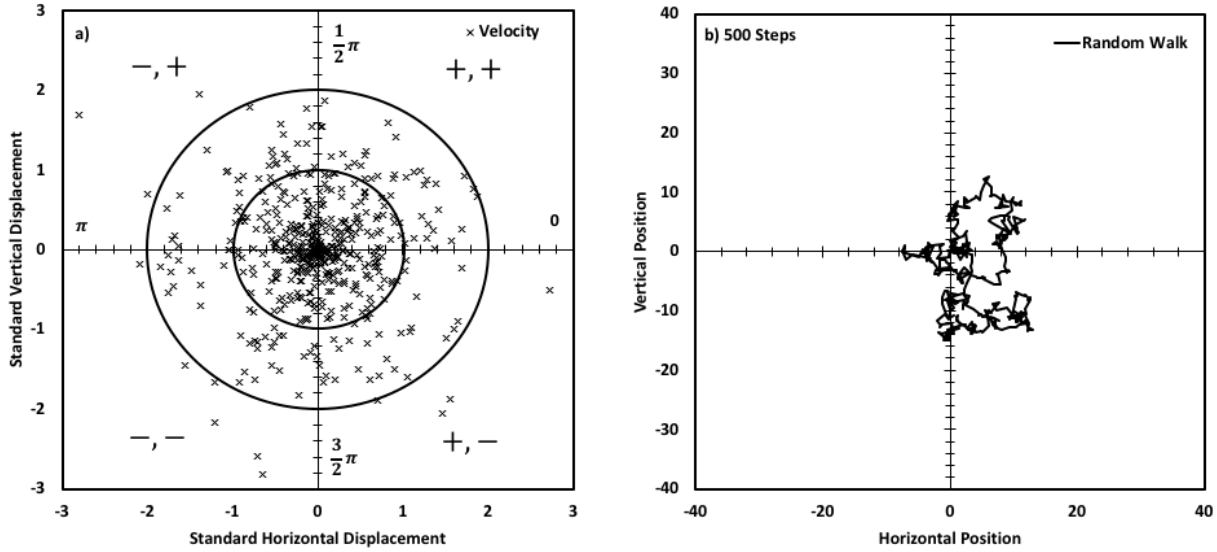

Figure 4.2: A standard normal random walk in two-dimensions. Figure 4.2.a represents example random velocities of molecular displacement (normal distribution) and direction (uniform distribution), and Figure 4.2.b represents the path of a 500-step random walk, where the randomly selected velocities are ordered in time and presented in sequence.

4.1.2.2. Radial Diffusion

Transforming the one-dimensional linear solution into polar coordinates rescales the linear displacement PDF to more accurately reflect Perrin's experimental results. To accomplish this, conservation of probability relates probabilistic diffusion for the linear and radial solutions. The next steps rely on $\int \mathcal{I}_{x,t,\mathcal{N}}\, dx = 1$ from Equation 4.1 to ensure that rescaling the distribution will conserve probability. However, the one-dimensional linear solution must be reimagined to characterize the probability over the interval of an angular increment $d\theta$. Assuming this reinterpretation also conserves probability, then $\iint \frac{dp^*_{\ell,t}}{dn} d\theta d\ell = 1$, where $d\theta d\ell = dx$ is an expression of an increment in the measurement space from Equation 4.1.

If there exists an analogous radial PDF for molecular displacement $\frac{dp^*_{r,t}}{dn}$ with unit area, $\int \frac{dp^*_{r,t}}{dn} dr = 1$, then these relationships can be equated through conservation of probability as $\int \frac{dp^*_{r,t}}{dn} dr = \iint \frac{dp^*_{\ell,t}}{dn} d\theta d\ell$. This equality allows us to infer a direct relationship between the linear and radial PDFs in one dimension as:

$$\because dr = d\ell\,, \qquad \therefore \frac{dp^*_{r,t}}{dn} = \int \frac{dp^*_{\ell,t}}{dn} d\theta \implies \frac{d^2p^*_{r,t}}{dnd\theta} = \frac{dp^*_{\ell,t}}{dn} \tag{4.5}$$

Where, $\frac{dp^*_{r,t}}{dn}$ is the zero-centered radial PDF that allows the random walk to travel in any direction while conserving probability. Substitution of the relationship $\frac{dp^*_{\ell,t}}{dn} = \frac{d^2p^*_{r,t}}{dnd\theta}$ into Equation 4.4 implies an analogous diffusion PDE, where probability is conserved in polar coordinates as:

$$\frac{d}{dt}\left[\frac{d^2p^*_{r,t}}{dnd\theta}\right] = D_{\ell,t,n}\frac{d^2}{d\ell^2}\left[\frac{d^2p^*_{r,t}}{dnd\theta}\right] \implies \frac{d^2p^*_{r,t}}{dnd\theta} = \frac{1}{\sqrt{s^2_{\ell,t,n}}}\bar{p}_z \tag{4.6}$$

Where, $\frac{d^2p^*_{r,t}}{dnd\theta}$ is the intensive property $\mathcal{I}_{x,t,\mathcal{N}}$ for the one-dimensional solution in polar coordinates. Although the PDE in Equation 4.6 still reflects a one-dimensional form of diffusion, it is a key intermediary step to interpret the solution to two-dimensional diffusion.

The probabilistic PDE from Equation 4.6 allows this analysis to reinterpret diffusion as a radial process. A single step within a random walk in polar coordinates is a scaled version of a single step from a binomial distribution. In polar coordinates, displacement can be both positive or negative with equal probability in all radial angles. Rotating the one-dimensional PDF by a half-circle from $0 \to \pi$, while conserving unit area properly scales the diffusion coefficient in polar coordinates. Evaluating the integral of the intensive property $\frac{d^2p^*_{r,t}}{dnd\theta}$ from Equation 4.6 with respect to $d\theta$ produces a zero-centered probabilistic PDF in polar coordinates $\frac{dp^*_{r,t}}{dn}$ as:

$$\frac{dp^*_{r,t}}{dn} = \frac{\int_0^\pi d\theta}{\sqrt{s^2_{\ell,t,n}}}\bar{p}_z \implies \frac{dp^*_{r,t}}{dn} = \frac{1}{\sqrt{\frac{s^2_{\ell,t,n}}{\pi^2}}}\bar{p}_z \tag{4.7}$$

The relationship in Equation 4.7 shows that the solution to the PDE in polar coordinates has the same form as Einstein with a scaled variance term. Assume there exists a radial variance such that Equation 4.7 is the solution to a second-order homogeneous PDE, then the linear and radial variance terms can be related as:

$$s_{r,t,n}^2 = \frac{s_{\ell,t,n}^2}{\pi^2} \implies \frac{dp_{r,t}^*}{dn} = \frac{1}{\sqrt{s_{r,t,n}^2}} \bar{p}_z \tag{4.8}$$

Where, $s_{r,t,n}^2$ represents the variance of displacement for one dimension in polar coordinates. By considering the Einstein-Smoluchowski equation in polar coordinates, this analysis can relate the radial diffusion coefficient $D_{r,t,n}$ to the linear diffusion coefficient $D_{\ell,t,n}$ from Einstein. Rescaling the variance into polar coordinates relates the linear and radial diffusion coefficients as:

$$s_{r,t,n}^2 = \frac{2}{\pi^2} D_{\ell,t,n} t, \qquad s_{r,t,n}^2 = 2D_{r,t,n} t \implies D_{r,t,n} = \frac{1}{\pi^2} D_{\ell,t,n} \tag{4.9}$$

Here, the radial diffusion coefficient is equated with the linear diffusion coefficient through a variance transformation, $s_{r,t,n}^2 = \frac{s_{\ell,t,n}^2}{\pi^2}$. This allows evaluation of the radial diffusion coefficient for a random walk in polar coordinates using direction-independent measures of displacement ℓ, while assuming radial symmetry over a consistent time interval.

This probabilistic interpretation of a random walk naturally breaks down when considering diffusion in two dimensions, because both the linear and radial diffusion coefficients are constrained to one dimension, as either r or ℓ. As soon as the random walk moves in two or more directions, the linear analogy is no longer appropriate and requires an area interpretation to conserve probability. Sequential steps in different directions implies that the position of the random walk will exist within some area relative to the origin. The next section creates an analogous PDF that conserves probability and implies a governing PDE in two dimensions.

4.1.2.3. Probabilistic Area-based Diffusion

The primary goal of this section is to generate a PDE that conserves probability for diffusion in two dimensions. Previously, this analysis demonstrates an analogous solution to Einstein's probabilistic PDE for one-dimensional diffusion in polar coordinates. However, this solution is a nonphysical representation of diffusion. The linear solution in polar coordinates applies to only the first step in the random walk before the origin resets to the center of the molecule (see Figure 4.2). Consider diffusion as a two-dimensional process and imagine a molecule tracing a path from an origin position to an end position (see Figure 4.2.b). The displacement of each

molecule can be either positive or negative at any angle between $0 \leq \theta \leq \pi$, in a two-dimensional plane, and hence generate a two-dimensional random walk. This section applies a unit circle interpretation to generate a probabilistic area and corresponding PDF that conserve probability for sequential steps in two dimensions. Starting from the foundation of Einstein's probabilistic PDE, the following progression will apply a transformation into two dimensions to characterize a physical representation of diffusion.

To begin, it is assumed that there exists some two-dimensional PDF $\frac{d^2 p^*_{\mathbb{A},t}}{dnd\theta}$ that is consistent with its one-dimensional counterpart $\frac{dp^*_{r,t}}{dn}$ that conserves probability for sequential steps in two-dimensions. Notably, the measurement space reflects area displacement in two dimensions, where $d\mathbb{A} \equiv dx$ for the general-form Equation 4.1. This implies a CDF that is a probabilistic volume $\mathbb{P}$, where the PDF $\frac{d^2 p^*_{\mathbb{A},t}}{dnd\theta}$ is analogous to height as:

$$\mathbb{P} = \int \frac{d^2 p^*_{\mathbb{A},t}}{dnd\theta} d\mathbb{A} \tag{4.10}$$

Where, $\mathbb{P}$ is a probabilistic volume of an area-based PDF for one angular interval about the origin and $\frac{d^2 p^*_{\mathbb{A},t}}{dnd\theta}$ is the intensive property $\mathcal{I}_{x,t,\mathcal{N}}$ for two-dimensional molecular displacement. Notably, this analysis defines the intensive property $\frac{d^2 p^*_{\mathbb{A},t}}{dnd\theta}$ to invoke conservation of probability through $\int \frac{d^2 p^*_{\mathbb{A},t}}{dnd\theta} d\mathbb{A} = 1$, where $dn \equiv d\mathcal{N}$ from Equation 4.1. To proceed, the analysis uses the area of a unit circle to equate the area-based PDF that corresponds to the radial diffusion PDF from Equation 4.8.

Here, the area-based process is reimagined through a probabilistic radius, $\mathbb{R} = \int \frac{dp^*_{r,t}}{dn} dr$, that approaches unity when evaluated over the entire measurement space: $\lim_{dr \to \infty} \mathbb{R} = 1$. By considering the area of a unit circle relative to the probabilistic radius R, this analysis can enforce conservation of probability for a two-dimensional process. Under these conditions, $\mathbb{P} = 2 \iiint d\mathbb{R}\, d\mathbb{R} d\theta$ represents the area of the unit circle. This clearly states a relationship between probabilistic radius and volume per angular increment as $\frac{d\mathbb{P}}{d\theta} = \mathbb{R}^2$. Upon substitution, the one- and two-dimensional diffusion PDFs can be related through the following:

$$\because \mathbb{R} = \int \frac{dp^*_{r,\mathrm{t}}}{dn} dr, \qquad \therefore \frac{d\mathbb{P}}{d\theta} = \left[\int \frac{dp^*_{r,\mathrm{t}}}{dn} dr\right]^2 \tag{4.11}$$

To proceed, the progression applies a change of variable, substitutes Equation 4.10 into Equation 4.11, and evaluates the area-based PDF $\frac{d^2 p^*_{\mathbb{A},t}}{dnd\theta}$. The probabilistic area is defined as $d\mathbb{A} = drdrd\theta$, which then transforms Equation 4.10 into $\frac{d\mathbb{P}}{d\theta} = \iint \frac{d^2 p^*_{\mathbb{A},t}}{dnd\theta} drdr$.

$$\because \frac{d\mathbb{P}}{d\theta} = \iint \frac{d^2 p^*_{\mathbb{A},t}}{dnd\theta} drdr, \qquad \therefore \iint \frac{d^2 p^*_{\mathbb{A},t}}{dnd\theta} drdr = \left[\int \frac{dp^*_{r,\mathrm{t}}}{dn} dr\right]^2 \tag{4.12}$$

By assuming that there is some governing PDE for the area-based PDF, then this progression can infer that $\frac{d^2 p^*_{\mathbb{A},t}}{dnd\theta}$ is an intensive property $\mathcal{I}_{x,t,\mathcal{N}}$ from Equations 4.1 and 4.2. Figure 4.3 shows that probabilistic radius $\mathbb{R} = \int \frac{dp^*_{r,t}}{dn} dr$ is equivalent to a radial CDF in one dimension.

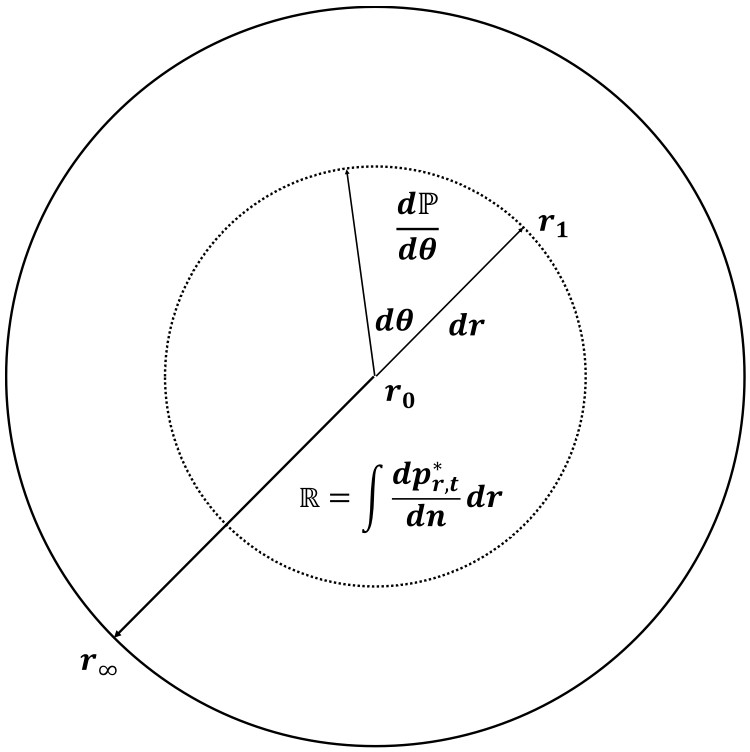

Figure 4.3: Geometric representation of area-based PDF.

As the radial increment increases from $r_0 \to r_1 \to r_\infty$, the probabilistic radius approaches unity as $\lim_{dr\to\infty} \mathbb{R} = 1$. Therefore, the area of the unit circle naturally conserves probability in two dimensions when evaluated indefinitely.

To proceed, the analysis can evaluate derivatives of the probabilistic volume per angular interval $\frac{d\mathbb{P}}{d\theta}$ with respect to radial increment dr to directly relate the PDFs for one- and two-dimensional diffusion. The resulting third-order derivative $\frac{d^3\mathbb{P}}{d\theta dr^2}$ provides an appropriate form to express the PDFs through $\frac{d\mathbb{R}}{dr} = \frac{dp^*_{r,t}}{dn} \iff \mathbb{R} = \int \frac{dp^*_{r,t}}{dn} dr$.

$$\frac{d\mathbb{P}}{d\theta} = \mathbb{R}^2 \implies \frac{d^2\mathbb{P}}{d\theta dr} = 2\frac{d\mathbb{R}}{dr}\mathbb{R} \implies \frac{d^3\mathbb{P}}{d\theta dr^2} = 2\frac{d}{dr}\left[\frac{d\mathbb{R}}{dr}\right]\mathbb{R} + 2\left[\frac{d\mathbb{R}}{dr}\right]^2 \tag{4.13}$$

By evaluating Equation 4.13 for a spatially-indefinite interval the probabilistic radius is constrained to unity as $\mathbb{R} = 1$. Upon substitution into Equation 4.13 and a derivation in Appendix C.1, this analysis demonstrates that $\frac{d^3\mathbb{P}}{d\theta dr dr} = 2\left[\frac{d\mathbb{R}}{dr}\right]^2$. This allows us to infer the following relationship:

$$\frac{d^2 p^*_{\mathbb{A},t}}{dn d\theta} = 2\left[\frac{dp^*_{r,t}}{dn}\right]^2 \tag{4.14}$$

Evaluating Equation 4.14 as the multiplication of two normal distributions $\left[\frac{dp^*_{r,t}}{dn}\right]^2 = \left[\frac{1}{\sqrt{s^2_{r,t,n}}}\bar{p}_z\right]^2$ allows us to generalize this relationship in Appendix C.2. Using a similar approach as Bromiley (2003) to multiply normal distributions, the resulting distribution for a two-dimensional process is transformed back into an analogous one-dimensional PDF with proper scale. The solution can be generalized for the area-based PDF as:

$$\frac{1}{2}\frac{d^2p^*_{\mathbb{A},t}}{dnd\theta} = \frac{1}{\sqrt{2\pi}s^2_{r,t,n}}\bar{p}_z \implies \frac{1}{2}\frac{d^2p^*_{\mathbb{A},t}}{dnd\theta} = \frac{1}{4\pi D_{r,t,n}t}\exp\left(-\frac{[r-m_r]^2}{4D_{r,t,n}t}\right)$$
$$\frac{1}{2}\int\frac{d^2p^*_{\mathbb{A},t}}{dnd\theta}dt = \int\frac{1}{4\pi D_{r,t,n}t}\exp\left(-\frac{[r-m_r]^2}{4D_{r,t,n}t}\right)dt \tag{4.15}$$

Clearly, the time-integral of the intensive property $\frac{d^2p^*_{\mathbb{A},t}}{dnd\theta}$ from Equation 4.15 is consistent with the solution to the Fourier equation for heat conduction in solids (Carslaw, 1921). In fact, this probabilistic interpretation suggests that the solution presented in Carslaw may be the time-integral of an area-based PDF. Temporal-continuity from Equation 4.1 provides the rational for taking the time-integral of Equation 4.15 as $\mathcal{E}_{x,\mathcal{N}} = \int \mathcal{I}_{x,t,\mathcal{N}}\,dt$.

Equation 4.15 represents a PDF solution for the area-based diffusion process as, $\frac{d^2p^*_{\mathbb{A},t}}{dnd\theta} = 2\left[\frac{1}{\sqrt{2\pi}s^2_{r,t,n}}\bar{p}_z\right]$. The PDF $\frac{d^2p^*_{\mathbb{A},t}}{dnd\theta}$ needs to be interpreted in one-dimensional polar coordinates to relate this solution to Einstein's one-dimensional diffusion process. Although this relationship is derived by scaling the standard normal PDF $\bar{p}_z$ to conserve probability in two dimensions, this solution can project into one dimension as:

$$\because \frac{d^2p^*_{\mathbb{A},t}}{dnd\theta} = \frac{2}{\sqrt{2\pi}s^2_{r,t,n}}\bar{p}_z, \qquad \therefore \frac{d^2p^*_{\mathbb{A},t}}{dnd\theta} = \left[\frac{1}{\sqrt{\frac{\pi}{2}}D_{r,t,n}t}\right]\bar{p}_z$$
$$Implies, \qquad \frac{d}{dt}\left[\frac{d^2p^*_{\mathbb{A},t}}{dnd\theta}\right] = D_{\mathbb{A},t,n}\frac{d^2}{dr^2}\left[\frac{d^2p^*_{\mathbb{A},t}}{dnd\theta}\right] \tag{4.16}$$

Where, the probabilistic solution to the PDE considers two-dimensional diffusion as an intensive process $\mathcal{I}_{x,t,\mathcal{N}} = \frac{d^2p^*_{\mathbb{A},t}}{dnd\theta}$. Reinterpreting this relationship as a solution to a second-order homogeneous PDE allows this analysis to infer an analogous variance measurement in two dimensions. Again, the Einstein-Smoluchowski equation is applied to develop a relationship between the area-based and radial diffusion coefficients.

$$s_{\mathbb{A},t,n}^2 = \sqrt{\frac{\pi}{2}} D_{r,t,n} t, \qquad s_{\mathbb{A},t,n}^2 = 2D_{\mathbb{A},t,n} t \implies D_{\mathbb{A},t,n} = \sqrt{\frac{\pi}{8}} D_{r,t,n} \tag{4.17}$$

The next section shows that Equation 4.16 leads directly to Fick's Law of diffusion after evaluating the spatial- and temporal-integrals of the intensive property $\frac{d^2 p_{\mathbb{A},t}^*}{dnd\theta}$. This interpretation provides a clear relationship between the one-dimensional linear, one-dimensional radial, and the area-based diffusion coefficients, which is shown to be consistent with the diffusion coefficient from Fick's Law. Appendix C.3 presents numerical examples of two-dimensional random walks and appropriately-scaled probabilistic solutions following from $D_{\mathbb{A},t,n} = \sqrt{\frac{\pi}{8}} D_{r,t,n}$.

4.1.3. FICK'S LAW

The area-based PDF $\frac{d^2 p_{\mathbb{A},t}^*}{dnd\theta}$ is an intensive property $\mathcal{I}_{x,t,\mathcal{N}}$. The extensive counterpart can be evaluated by integrating the intensive property with respect to the number of measurements as $\mathcal{E}_{x,t} = \int \mathcal{I}_{x,t,\mathcal{N}}\, d\mathcal{N}$. For molecular diffusion, the number of measurements is equal to the number of molecules as $n \equiv \mathcal{N}$. Upon integration of the PDF from Equation 4.16, the extent of displaced molecules can be estimated as: $\frac{d^2 p_{\mathbb{A},t}^*}{d\theta} = \int \left[\frac{1}{\sqrt{\frac{\pi}{2}} D_{r,t,n} t} \bar{p}_z \right] dn$. In the context of diffusion, the new relationship reflects molecular displacement for some source λ as:

$$\frac{dp_{\mathbb{A},t}^*}{d\theta} = \frac{\int_0^\lambda dn}{\sqrt{s_{\mathbb{P},t,n}^2}} \bar{p}_z \implies \frac{dp_{\mathbb{A},t}^*}{d\theta} = \left[\frac{\lambda}{\sqrt{\frac{\pi}{2}} D_{r,t,n} t} \right] \bar{p}_z \tag{4.18}$$

Where, $\frac{dp_{\mathbb{A},t}^*}{d\theta}$ is an extensive PDF per angular increment with the molecular displacement of source λ resulting from diffusion after time t. Integrating the PDF in Equation 4.18 by the measurement space $d\mathbb{A} \equiv dx$, estimates the total molecular displacement from the origin. In this context, the extensive PDF $\mathcal{E}_{x,t} = \frac{dp_{\mathbb{A},t}^*}{d\theta}$ estimates the frequency of displacement related to source λ and has units of length per angular increment. This interpretation provides a mathematical expression for the expected number of moles within probabilistic area $\mathbb{A}$ due to diffusion after time t. The concept

of intensive and extensive properties provides a link between the probabilistic solution to Brownian motion and concentration-based diffusion according to Fick's law.

Here, integrating the diffusion PDF $p^*_{\mathbb{A},t}$ with respect to intervals of time dt and probabilistic area $d\mathbb{A}$ represents the probability that a molecule exists within some predefined volume V. Dividing the number of moles by the predefined volume derives the composite property of concentration as:

$$C = \frac{n}{V} \implies C = \frac{\lambda}{V}\iint p^*_{\mathbb{A},t}\, d\mathbb{A}dt \implies \frac{d}{dt}\left[\frac{dC}{d\mathbb{A}}\right] = \frac{\lambda}{V}p^*_{\mathbb{A},t} \tag{4.19}$$

Where, $\int p^*_{\mathbb{A},t}\, d\mathbb{A}$ represents the probability that molecules from source λ are inside the area increment $d\mathbb{A}$. This relationship is a unique interpretation of concentration that relies upon both Einsteinian diffusion and the solution to the Fourier heat transfer equation from Equation 4.15. Integrating Equation 4.18 by angular increment $d\theta$ and evaluating for a symmetric physical process over interval $0 \le \theta \le \pi$ defines how the concentration per probabilistic area $\frac{d^2C}{d\mathbb{A}dt}$ will change with respect to time.

$$p^*_{\mathbb{A},n} = \frac{\lambda\int_0^\pi d\theta}{\sqrt{\frac{\pi}{2}D_{r,t,n}t}}\bar{p}_z \implies \frac{d}{dt}\left[\frac{dC}{d\mathbb{A}}\right] = \frac{1}{\sqrt{\frac{1}{2\pi}\frac{V}{\lambda}D_{r,t,n}t}}\bar{p}_z \tag{4.20}$$

Note that $\frac{d}{dt}\left[\frac{dC}{d\mathbb{A}}\right]$ can be expanded on the basis that $\frac{dC}{d\mathbb{A}} = \frac{d}{d\mathbb{A}}\left[\frac{dn}{dV}\right] = \frac{d^2n}{d\mathbb{A}dV}$. Therefore, $\frac{d}{dt}\left[\frac{dC}{d\mathbb{A}}\right] = \frac{d}{dt}\left[\frac{d^2n}{d\mathbb{A}dV}\right]$ is a representation of a PDF where $\frac{d^2n}{d\mathbb{A}dV}$ is an intensive measurement $\mathcal{I}_{x,t,\mathcal{N}}$. Under ideal conditions, the diffusion coefficient $D_{r,t,n}$, source strength λ, and volume V are constants, which constrains the time-derivative to express a normal distribution. Under these circumstances, Equation 4.20 conforms to Fick's Law PDE for a predefined probabilistic area $\mathbb{A}$, and relate its diffusion coefficient $D_{n,t,V}$ to the radial diffusion coefficient $D_{r,t,n}$ through the Einstein-Smoluchowski equation. Given a predefined probabilistic area $\mathbb{A}$, the number of moles n and the volume V characterizes the diffusion coefficient $D_{n,t,V}$ with respect to concentration.

$$\frac{d}{dt}\left[\frac{d^2n}{d\mathbb{A}dV}\right] = D_{n,t,V}\frac{d^2}{dr^2}\left[\frac{d^2n}{d\mathbb{A}dV}\right]$$

$$\sqrt{\frac{1}{2\pi}}\frac{V}{\lambda}D_{r,t,n}t = 2D_{n,t,V}t \implies D_{n,t,V} = \sqrt{\frac{1}{2\pi}}\frac{V}{\lambda}D_{r,t,n} \tag{4.21}$$

This interpretation of Fick's Law shows that $\frac{d^2n}{d\mathbb{A}dV}$ integrates to an extensive property through $\mathcal{E}_{x,t} = \int \frac{d^2n}{d\mathbb{A}dV}\,dV$. Note that $D_{n,t,V}$ is inversely proportional to the molar density $\frac{\lambda}{V}\left[\frac{moles}{m^3}\right]$ consistent with the observation that the diffusion coefficient for gases is greater than that of liquids. The outcome of this derivation is that the general-form PDE from Equation 4.2 characterizes both one- and two-dimensional diffusion as well as Fick's Law of diffusion. Table 4.2 summarizes the analogies for the linear diffusion coefficient through the radial diffusion coefficient, the Fourier equation analogy using an area-based diffusion coefficient, and finally to Fick's Law. Moreover, it is clear that diffusive mass transport $D_{n,t,V}$ across space is entirely a function of the Brownian motion coefficient $D_{\ell,t,n}$ that proceeds independent of any measurable ambient process except time. Hence, Brownian motion epitomizes the essence of spatial- and temporal-continuity as discussed in Equation 4.1.

Table 4.2: Scaled diffusion coefficients and PDEs for various dimensionality.

Dimensionality	Probabilistic PDE	Diffusion Coefficient
Linear-Magnitude (Einstein)	$\frac{d}{dt}\left[\frac{dp^*_{\ell,t}}{dn}\right] = D_{\ell,t,n}\frac{d^2}{d\ell^2}\left[\frac{dp^*_{\ell,t}}{dn}\right]$	$D_{\ell,t,n} = \frac{s^2_{\ell,t,n}}{2t}$
Linear-Polar (Perrin)	$\frac{d}{dt}\left[\frac{dp^*_{r,t}}{dn}\right] = D_{r,t,n}\frac{d^2}{dr^2}\left[\frac{dp^*_{r,t}}{dn}\right]$	$D_{r,t,n} = \frac{D_{\ell,t,n}}{\pi^2}$
Two-dimensional (Fourier)	$\frac{d}{dt}\left[\frac{dp^*_{\mathbb{A},t}}{dn}\right] = D_{r,t,n}\frac{d^2}{dr^2}\left[\frac{dp^*_{\mathbb{A},t}}{dn}\right]$	$D_{\mathbb{A},t,n} = \sqrt{\frac{\pi}{8}}D_{r,t,n}$
Fick's Law	$\frac{d}{dt}\left[\frac{d^2n}{d\mathbb{A}dV}\right] = D_{n,t,V}\frac{d^2}{dr^2}\left[\frac{d^2n}{d\mathbb{A}dV}\right]$	$D_{n,t,V} = \sqrt{\frac{1}{2\pi}}\frac{V}{\lambda}D_{r,t,n}$

4.2. DISCUSSION

The theory developed in this chapter attempts to bridge the gap between experimental results to develop governing PDEs that reproduce measurement distributions under non-ideal conditions that ensure conservation of probability, as well as spatial- and temporal-continuity. Chapter 3 shows that a general-form PDE produces parametric PDFs that are consistent with observation data and do not conform to a scaled normal distribution. There are some analogies between the water consumption and physical applications that suggest a broader generalization. These applications have temporal and spatial derivatives that are related through scaling coefficients. It is suspected that the scaling coefficients have a very specific meaning within a transient system. Here, this chapter contends that there are three general concepts which govern transient probabilistic systems: 1) a source/sink term, 2) the measured property, and 3) a conduit that connects the source/sink to the measured property. For the water consumption application, the necessity of water to maintain standard of living conditions is the source term, household-specific qualities that compel water consumption represents the conduit, and water consumption is the measured property.

The conduit is the most interesting aspect of this metaphor because it connects the source/sink term to the measured property. In many cases, the properties of the conduit may be difficult to measure directly and may vary with time or system conditions. For instance, the conduit for water consumption may be the resistance of market participants to changing consumption habits as factors like price and weather change. Ultimately, pairs of source/sink terms and the measured property may provide the opportunity to quantify the conduit and even how it changes through time. This indirect evaluation may provide a predictive model to anticipate the resulting measured property for different source/sink terms. The water consumption analysis in Chapter 3 is an application of this approach and motivates this discussion of the philosophical implications of quantifying the conduit.

The virtue of the water consumption example from Figure 4.1 is that it exemplifies a spatial distribution for a specific interval of time. By evaluating sequential temporal groupings, the approach is able to infer the time-dependent evolution of the spatial distribution of measurements. The water consumption application required multiple temporal groupings of data to infer causality between ambient processes and the time-dependent evolution of the system. Here, the relationship

between the spatial- and temporal-evolution of the intensive property $\mathcal{I}_{x,t,\mathcal{N}}$ can be expressed using a PDE in the context of a Taylor expansion as:

$$\underbrace{\frac{d}{dt}\left[\mathcal{I}_{x,t,\mathcal{N}}\right]}_{temporal\ distribution} = \underbrace{\frac{\partial x}{\partial t}\frac{\partial}{\partial x}\left[\mathcal{I}_{x,t,\mathcal{N}}\right] + \frac{\partial^2 x}{\partial t^2}\frac{\partial^2}{\partial x^2}\left[\mathcal{I}_{x,t,\mathcal{N}}\right]dt + \cdots}_{\substack{spatial\ distribution \\ with \\ scaling\ coefficients}} \tag{4.27}$$

The PDE in Equation 4.27 is a general representation of the relationship between space, time, and the conduit that provides a medium for changes to the intensive property. Notably, additional terms beyond a homogeneous second-order PDE allow for a PDF solution $\mathcal{I}_{x,t,\mathcal{N}}$ that does not necessarily conform to a normal distribution. Equation 4.27 shows that the intensive PDF $\mathcal{I}_{x,t,\mathcal{N}}$ is simultaneously a distribution in both space and time. The case of a second-order homogeneous PDE, which generates a normal distribution for $\mathcal{I}_{x,t,\mathcal{N}}$, relates to a specific parameterization of Equation 4.27, where $\frac{\partial x}{\partial t} = 0, \frac{\partial^3 x}{\partial t^3} = 0, \ldots$ and $D_{x,\mathcal{N},t} = \frac{\partial^2 x}{\partial t^2}dt$ to produce $\frac{d}{dt}\left[\mathcal{I}_{x,t,\mathcal{N}}\right] = D_{x,\mathcal{N},t}\frac{d^2}{dx^2}\left[\mathcal{I}_{x,t,\mathcal{N}}\right]$. For the diffusion process, the molecules themselves represent the conduit of displacement through molecular interactions. Notably, the self-diffusion coefficient from Einstein of the form $D_{x,\mathcal{N},t}$ correlates with the size and shape of molecules that comprise the liquid. Likewise, this analogy allows for characterization of the conduit that connects the ambient processes of price and weather to water consumption.

The water consumption application provides parameterization values for the right-hand side of Equation 4.27 for the specific influence of price and weather on the statistics that describe the water consumption PDF. Parameters within the water consumption PDE are an expression of consumer preference that result from the water consumption response to changes in price and weather. The high level of accuracy in reproducing the evolution of the water consumption PDF suggests that this approach could be extended as a forecasting tool, while considering additional ambient processes beyond price and weather. Although this parameterization potentially represents an over-simplification of a complex issue, the parameters of the water consumption PDE appear to capture the evolution of consumer preference. This theory is predicated on the relative stability of microscopic processes that contribute to water consumption. Chapter 3 suggests that these processes include number of occupants, household and yard sizes, and

household income, among others. Assuming that these water consumption characteristics remain relatively stable, any changes to the water consumption PDF must be attributed to macroscopic, ambient conditions. Future work might parameterize a similar water consumption PDE for a different municipality and compare them to see if the parameterization, as a representation of consumer preference, is sensitive to these city-specific conditions.

4.3. CONCLUSIONS

This chapter applies conservation of probability as an all-encompassing conservation law that is demonstrated as a solution that applies to both physical and abstract processes. The water consumption application demonstrates the importance of collecting relative frequency measurements when quantifying the evolution of complicated systems dynamics. The outcome of this chapter presents a fundamental relationship between viewing observations as frequency histograms, and their associated PDFs, as the solution to governing parametric PDEs.

This chapter demonstrates the importance of using dimensionally consistent scaling of a normal distribution to express diffusion as a two-dimensional process. The area of the unit circle provides the basis for transforming Einstein's one-dimensional diffusion PDE into an area-based diffusion process. Furthermore, this solution is probabilistic representation that results in a solution that is consistent with that of the Fourier equation for heat conduction in solids. Subsequent spatial- and temporal-integration of the two-dimensional diffusion process yields a probabilistic derivation that is consistent with Fick's Law.

Finally, the generality of this approach to govern both abstract and physical processes allows this analysis to conclude that second-order PDEs reflect interactions between three well-defined system components: 1) a source/sink term, 2) the measured property, and 3) a conduit that connects the source/sink term to the measured property. This realization motivates further investigation into how parameterization of a governing PDE can meaningfully describe and quantify the properties of the conduit. In conclusion, transient processes are shown to be characterized by the general flow of information, which can be used to describe both physical and abstract processes.

5. Thesis Conclusions

Through automation and the advancements of information science, there exist large quantities of data that researches use to describe and quantify the world around them. This thesis considers measurement data from sources as disparate as economics, science and engineering, stock market indices, and digital photo processing, and demonstrates a general-form methodology for expressing these datasets as continuous probability density functions (PDFs). Moreover, this thesis demonstrates that asymmetric, shifted, heavy-tailed, albeit continuous PDFs are the solution to an "advective-dispersive" like transport process resulting from a governing parametric partial differential equation (PDE). Each technical chapter considers the following three foundational interpretations of measurement data histograms that allow for a smooth transition between steady-state conditions through the transport process. These foundational themes are: 1) the hierarchal relationship between the parametric control function, standard-score PDF, and standard-score cumulative distribution function (CDF); 2) interpreting the measurement space PDF as the solution to a transport process, where the median represents advection and a combination of the standard deviation and standard-score PDF represent dispersion; and, 3) the relationship between extensive measurements and their corresponding intensive PDF; conservation of probability through the measurement space CDF; and the notion of spatial and temporal-continuity of measurement PDFs that result in a continuum representation of otherwise discrete data histograms. Clearly, conservation of probability is the predominant theme of this thesis and relates the three technical chapters through the standard-score PDF, evolution of the measurement space PDF, and finally the transient nature of the measurement space CDF. The foundational themes are demonstrated via application of residential water consumption, with the intent of providing water utilities with a methodology for parameterizing PDEs that can then be used to forecast transient residential water consumption PDFs under the influence of ambient processes of price and weather.

The conclusion of this thesis is that conservation of probability appears to be an all-encompassing conservation law for measurement information and may assist in describing the evolution of transient continuum systems. In this context, probability is used to relate frequency as being a relative measurement of information, where this information constitutes a set of measurements recording the instantaneous state of both physical and abstract systems alike. Specifically, this thesis demonstrates a relationship between measurement histograms, parametric

PDFs, and PDEs for the physical process of molecular diffusion and the abstract economic system of water consumption. The outcome of this work is to motivate further discussion of parametric "advective-dispersive" like transport PDEs as being a more general expression of a systems continuum response, relative to the established notion of second-order homogeneous PDEs in the context of Brownian motion. These parametric PDEs reflect interactions between three well-defined system components: 1) a source/sink term, 2) the measured property, and 3) a conduit that connects the source/sink to the measured property. Finally, there is compelling evidence that parameters within the governing PDE meaningfully describe and quantify the properties of the conduit, much the same as the diffusion coefficient in the second-order homogeneous PDE representing Brownian motion is essentially a second-order partial derivative arising from a Taylor expansion. Ultimately, conservation of probability provides a mechanism for reconciliation that ensures no information is either created or destroyed, while generating PDEs to reproduce spatially- and temporally-continuous PDF solutions that coincide with measurement data histograms.

REFERENCES

American Water Works Association (AWWA), 2012. Buried No Longer: Confronting America's Water Infrastructure Challenge. http://www.awwa.org/legislation-regulation/issues/infrastructure-financing.aspx. Accessed 2/26/2015.

Beecher, J. & Chestnutt, T., 2010. "The Conservation Conundrum: How Declining Demand Affects Water Utilities." Journal American Water Works Association 102 (2):78.

Beecher, J., 2012. Declining Water Sales and Utility Revenues: A Framework for Understanding and Adapting. National Water Rates Summer: White Paper. Racine, Wisconsin. August 29-30, 2012.

Boland, J.J.; Ziegielewski, B.; Baumann, D.D.; and Optiz, E.M., 1984. Influence of Price and Rate Structures on Municipal and Industrial Water Use. US Corps of Engineers Institute for Water Resources, Fort Belvoir, Virginia 22060, Contract Report 84-C-2.

Bromiley, P. A., 2003. Products and Convolutions of Gaussian Probability Density Functions. Internal Memo – Tina Memo No. 2003-003. Last Updated 2014. Imaging Sciences Research Group, Institute of Population Health, School of Medicine, University of Manchester.

Brookeshire, D.S.; Burness, H.S.; Chermak, J.M.; and Krause, K., 2002. Western Urban Water Demand. Journal of Natural Resources, 42:873.

Carslaw, H. S., 1921. Introduction to the mathematical theory of the conduction of heat in solids. 2d ed., completely rev. London: Macmillan.

House-Peters, L. A. and Chang, H., 2011. Urban water demand modeling: Review of concepts, methods, and organizing principles. Water Resources Research, 47, 15.

Hutchinson, J., 2001. Culture, Communication, and an Information Age Madonna. IEEE Professional Communication Society Newsletter. May/June 2001 vol. 45 no. 3.

Hyndman, R.J.; Bashtannyk, D.M.; and Grunwald, G.K., 1996. Estimating and Visualizing Conditional Densities. Journal of Computational and Graphical Statistics, vol. 5, no. 4 pp. 315-336.

Dalhuisen, J.M.; Florax, R.J.G.M; de Groot, H.L.F.M; and Nijkamp, P., 2003. Price and Income Elasticities of Residential Water Demand. Tinbergen Institute Discussion Paper, TI 2003-057/3.

Donkor, E.A.; Mazzuchi, T.A.; Soyer, R.; and Roberson, J.A., 2011. Urban Water Demand Forecasting: Review of Methods and Models. Journal of Water Resour. Plann. Manage., 2014, 140(2): 146-159

J. Duda, 2017. Rapid parametric density estimation. arXiv:1702.02144v2 [cs.LG] 20 Feb 2017.

Einstein, A., 1905. Investigations on the Theory of, the Brownian Movement (A. D. Cowper, Trans.). Dover Publications, Inc.

Environment Canada, 2017. Kitchener/Waterloo Weather Station Historical Data. Accessed June 2016. http://climate.weather.gc.ca/historical_data/search_historic_data_e.html.

Environmental Protection Agency (EPA), 2003. Water and Wastewater Pricing – An Informational Overview. Office of Wastewater Management. EPA 832-F-03-027.

EPA, 2005. Case Studies of Sustainable Water and Wastewater Pricing. Office of Water. EPA 816-R-05-007.

EPA, 2006. Expert Workshop on Full Cost Pricing of Water and Wastewater Service. November 1-3, 2006. Michigan State University, Institute for Public Utilities. Office of Water. EPA 816-R-07-005.

Eskaf, S.; Hughes, J.; Tiger, M.; Bradshaw, K., & Leurig, S., 2014. "Measuring & Mitigating Water Revenue Variability: Understanding How Pricing Can Advance Conservation Without Undermining Utilities' Revenue Goals. UNC Environmental Finance Center and Ceres.

Espey, M.; Espey, J.; & Shaw, W.D., 1997. Price Elasticity of Residential Demand for Water: A Meta-analysis. Water Resources Research, 33:6:1369.

Gauss, C. F., 1809. Theoria Motus Corporum Celestium. Hamburg, Perthes et Besser. Translated as Theory of Motion of the Heavenly Bodies Moving about the Sun in Conic Sections (trans. C. H. Davis), Boston, Little, Brown 1857. Reprinted: New York, Dover 1963.

Gauss, C. F., 1816. Bestimmung der Genauigkeit der Beobachtungen. Zeitschrift für Astronomie und verwandte Wissenschaften. 1: 187–197.

Hughes, J. A. and Leurig, S., 2013. Assessing Water System Revenue Risk: Considerations for Market Analysts. University of North Carolina – Environmental Finance Center and Ceres.

Hunter, M.; Donmoyer, K.; Chelius, J.; & Naumick, G., 2011. "Declining Water Use Presents Challenges, Opportunities." Journal of American Water Works Association. 37:5:18.

Loáiciga, H.A., 2009. Derivation approaches for the Theis equation. Groundwater, 47(4), 1-4.

Lohman, S.W., 1972. Ground-water hydraulics, U.S. Geological Survey Prof. Paper 708, 70p.

Ministry of the Environment, 2011. Water Opportunities and Water Conservation Act, 2010. Ministry of the Environment, Ontario. Environmental Registry Number 010e9940, Retrieved from: http://www.e-laws.gov.on.ca/html/source/statutes/english/2010/elaws_src_s10019_e.htm. August 12, 2011.

Mayer, P.; DeOreo, W.; Chesnutt T.; & Summers, L., 2008. Water Budgets and Rate Structures – Innovative Management Tools. Journal of American Water Works Association, 100:5:117.

Mehan III, T. & Kline, I., 2012. "Pricing as a Demand-Side Management Tool: Implications for Water Policy and Governance." Journal American Water Works Association, 104:2:61.

Ministry of the Environment (MOE), 2011. Water Opportunities and Water Conservation Act, 2010. Ministry of the Environment, Ontario. Environmental Registry Number 010e9940, http://www.e-laws.gov.on.ca/html/source/statutes/english/2010/elaws_src_s10019_e.htm.

National Aeronautics and Space Administration (NASA). NASA Langley Research Center POWER Project. Latitude: 43.464, Longitude: 80.52. Accessed June 2016. https://power.larc.nasa.gov/cgi-bin/agro.cgi

Narasimhan, T. N., 1999. Fourier's Heat Conduction Equation: History, Influence, and Connections. Reviews of Geophysics 37, February 1999. Pages 151-172. American Geophysical Union. 8755-1209/99/1998RG900006515.

Olmstead, S.M.; Hanemann, W.M.; & Stavins, R.N., 2007. Water Demand Under Alternative Price Structures. Journal of Environmental Economics and Management, 54:2:181.

Pearson, K., 1894. On the dissection of asymmetrical frequency curves. Philosophical Transactions of the Royal Society A. 185: 71–110. doi:10.1098/rsta.1894.0003.

Perrin, J., 1913. Atoms (D. L. Hammick, Trans.). New York. Van Norstrand Company.

Region of Waterloo. Transportation and Environmental Services – Water Services. (2006a). Water Efficiency Master Plan Update 2007-2015. Available Online. http://www.regionofwaterloo.ca/en/aboutTheEnvironment/resources/water%20efficiency%20master%20plan.pdf

Region of Waterloo, 2006b. Environews. April, 2006. 8 pages. Available Online. Accessed July 2017. http://www.regionofwaterloo.ca/en/aboutTheEnvironment/resources/Spring06.pdf

Region of Waterloo – Department of Water Efficiency. Water Efficiency Master Plan 2015-2025. Available Online. http://www.regionofwaterloo.ca/en/aboutTheEnvironment/resources/WEMP_2015_-_2025_FINAL__May_1_2025.pdf

Rehan, R.; Knight, M.A.; Haas, C.T.; & Unger, A.J.A., 2011. Application of System Dynamics for Developing Financially Self-Sustaining Management Policies for Water and Wastewater Systems. Water Research, 45:16:4737.

Rehan, R.; Knight, M.A.; Unger, A.J.A.; & Haas, C.T., 2013. Development of a System Dynamics Model for Financially Sustainable Management of Municipal Watermain Networks. Water Research, 47:20:7184.

Rehan, R.; Knight, M.A.; Unger, A.J.A.; & Haas, C.T., 2014a. Financially Sustainable Management Strategies for Urban Wastewater Collection Infrastructure – Development of a System Dynamics Model. Tunnelling and Underground Space Technology, 39:116.

Rehan, R.; Unger, A.J.A.; Knight, M.A.; & Haas, C.T., 2014b. Financially Sustainable Management Strategies for Urban Wastewater Collection Infrastructure – Implementation of a System Dynamics Model. Tunnelling and Underground Space Technology, 39:102.

Rehan, R.; Unger, A.J.A; Knight, M.A.; & Haas, C.T., 2014c. Strategic Water Utility Management and Financial Planning Using a New System Dynamics Tool. Journal of American Water Works Association, 107:1.

Stahl, S., 2006. The evolution of the normal distribution. Mathematics Magazine.79(2), 96-113.

Stone, C., 1994. The Use of Polynomial Splines and Their Tensor Products in Multivariate Function Estimation. The Annals of Statistics, vol. 22, no. 1, pp. 118-171.

Stockwiz, 2009. Historical Data for S&P 500 Stocks. Retrieved from: http://pages.swcp.com/stocks/ July 2016.

Sudicky, E., 1986. A natural gradient experiment on solute transport in a sand aquifer: Spatial variability of hydraulic conductivity and its role in the dispersion process. Water Resources Research, 22(13), 2069–2082, DOI: 10.1029/WR022i013p02069

University of Waterloo. University of Waterloo Weather Station Historical Data. http://weather.uwaterloo.ca/data.html. Accessed June 2016.

USEPA, 2003. Water and wastewater pricing: An Informational Overview. Office of Wastewater Management. EPA 832-F-03-027.

USEPA, 2005. Case Studies of Sustainable Water and Wastewater Pricing. Office of Water. EPA 816-R-05-007. December 2005.

USEPA, 2006. Expert Workshop on Full Cost Pricing of Water and Wastewater Services. Final Summary Report. Michigan State University Institute for Public Utilities. November, 2006.

Worthington, A.C. & Hoffman, M., 2008. An Empirical Survey of Residential Water Demand Modelling. Journal of Economic Survey, 22:5:842.

Zambom and Dias 2012. A Review of Kernel Density Estimation with Applications to Econometrics. Universidade Estadual de Campinas. December 2012.

APPENDIX A.1

The derivation of the general-form PDF p_z begins with the characteristic ODE and rearranges this relationship to isolate the dependent of the PDF on the indefinite integral of the control function with respect to the standard-score variable z:

$$\frac{dp_z}{dz} = g_z p_z$$

$$\frac{1}{p_z}\frac{dp_z}{dz} = g_z \implies \int \frac{1}{p_z} dp_z = \int g_z \, dz$$

$$\ln|p_z| + C_1 = \int g_z \, dz \implies exp[ln|p_z| + C_1] = exp\left[\int g_z \, dz\right]$$

$$p_z = \frac{1}{C_2} exp\left[\int g_z \, dz\right], \qquad C_2 = \frac{1}{exp[C_1]}$$

$$p_z = exp\left[\int g_z \, dz\right] \; for \; C_2 = 1$$

Notably, the constant of integration C_2 from $\int \frac{1}{p_z} dp_z$ will be absorbed by the constant of integration from $\int g_z \, dz$ to ensure the PDF p_z reflects unit area on a definite interval within the standard-score space z. Therefore, assume that $C_1 = 0$ and $C_2 = 1$ to simplify this expression.

APPENDIX A.2

Assume a linear control function

$$g(z) = -(\alpha_1 + \alpha_2 z)$$

$$\int g(z)\,dz = -\int (\alpha_1 + \alpha_2 z)\,dz$$

$$f(z) = \exp(-\alpha_1 - \alpha_2 z - C), \qquad assume\ \alpha_0 = C$$

$$f(z) = e^{-\alpha_0} \exp\left(-\alpha_1 z - \frac{\alpha_2}{2} z^2\right)$$

Assume symmetry about $z = 0 \rightarrow \alpha_1 = 0$:

$$f(z) = e^{-\alpha_0} \exp\left(-\frac{\alpha_2}{2} z^2\right)$$

Median-centered distribution implies $\int_0^\infty f(z)\,dz \equiv \frac{1}{2}$

$$\int_0^\infty \left[e^{-\alpha_0} \exp\left(-\frac{\alpha_2}{2} z^2\right)\right] dz = \frac{1}{2}$$

$$e^{-\alpha_0} \int_0^\infty \left[\exp\left(-\frac{\alpha_2}{2} z^2\right)\right] dz = \frac{1}{2}$$

$$e^{-\alpha_0} \frac{\sqrt{\pi}}{2\sqrt{\frac{\alpha_2}{2}}} \operatorname{erf}\left(\sqrt{\frac{\alpha_2}{2}} z\right)\Bigg|_0^\infty = \frac{1}{2}$$

$$e^{-\alpha_0} \frac{\sqrt{2\pi}}{2\sqrt{\alpha_2}} (1 - 0) = \frac{1}{2}$$

$$e^{-\alpha_0} = \sqrt{\frac{\alpha_2}{2\pi}}$$

$$\alpha_0 = -\ln\left|\sqrt{\frac{\alpha_2}{2\pi}}\right|$$

Consider $\alpha_2 = 1, \alpha_0 = \ln\left|\sqrt{2\pi}\right|$

$$\text{Normal Distribution,} \qquad f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} z^2\right)$$

APPENDIX A.3

Assume the median $m_{x,i}$ and standard deviation $\sigma_{x,i}$ are known statistics of the measurement data x_i. From Table 2.1, the continuous measurement x can be related to the median-relative y and standard-score variable z as:

$$x = m_{x,i}y = m_{x,i} + \sigma_{x,i}z$$

Multiplying this relationship by the PDF p_x and integrating over some definite interval in the measurement space:

$$\int_{x_0}^{x_1} xp_x\,dx = m_{x,i}\int_{x_0}^{x_1} yp_x\,dx = m_{x,i}\int_{x_0}^{x_1} p_x\,dx + \sigma_{x,i}\int_{x_0}^{x_1} zp_x\,dx$$

Next, transform this relationship by $p_x dx = p_y dy = p_z dz$ from Table 2.1:

$$\int_{x_0}^{x_1} xp_x\,dx = m_{x,i}\int_{x_0}^{x_1} yp_x\,dy = m_{x,i}\int_{z_0}^{z_1} p_z\,dz + \sigma_{x,i}\int_{z_0}^{z_1} zp_z\,dz$$

Evaluating the indefinite interval of the PDF to be $\int_{z_0}^{z_1} p_z\,dz = 1$:

$$\int_{x_0}^{x_1} xp_x\,dx = m_{x,i}\int_{x_0}^{x_1} yp_x\,dy = m_{x,i} + \sigma_{x,i}\int_{z_0}^{z_1} zp_z\,dz$$

Finally, substituting the probability-weighted mean from Equation 2.11:

$$\int_{x_0}^{x_1} xp_x\,dx = m_{x,i}\int_{x_0}^{x_1} yp_x\,dy = m_{x,i} + \sigma_{x,i}\mu_z$$

Following from our understanding that the mean statistic occupies a position on the PDF within each spatial representation, this progression applies the notation from Equation 2.11 for the measurement and median-relative spaces:

$$\mu_x = \int_{x_0}^{x_1} xp_x\,dx, \qquad \mu_y = \int_{x_0}^{x_1} yp_x\,dy$$

which produces the following equalities:

$$\mu_x = m_{x,i}\mu_y = m_{x,i} + \sigma_{x,i}\mu_z$$

$$\mu_x = m_{x,i} + \sigma_{x,i}\int_{z_0}^{z_1} zp_z\,dz$$

$$\mu_x = m_{x,i} + \sigma_{x,i}\int_{z_0}^{z_1}\left[z\exp\left(\int g_z\,dz\right)\right]dz$$

APPENDIX A.4

Table A.4.1: Binned histogram data for the water demand, hydraulic conductivity, and S&P 500 index CDFs.

Median-Relative Bins y_k [−]	**Water Consumption** $c_{y,k}$ [−]	**Hydraulic Conductivity** $c_{y,k}$ [−]	**S&P 500 Index 08/21/2009** $c_{y,k}$ [−]
0.00	0.0000	0.0000	0.0000
0.25	0.0299	0.0264	0.0489
0.50	0.1327	0.0903	0.1813
0.75	0.2955	0.2806	0.3320
1.00	0.4900	0.5000	0.4990
1.25	0.6558	0.6764	0.6477
1.50	0.7763	0.7708	0.7495
1.75	0.8561	0.8528	0.8411
2.00	0.9109	0.9083	0.9104
2.25	0.9430	0.9597	0.9409
2.50	0.9631	0.9903	0.9715
2.75	0.9752	0.9944	0.9817
3.00	0.9839	0.9958	0.9939
3.25	0.9900	0.9986	0.9980
3.50	0.9944	1.0000	0.9980
3.75	0.9974	1.0000	0.9980
4.00	1.0000	1.0000	1.0000

Table A.4.2: Binned data for the Lenna light intensity CDF.

$x_k = m_{x,i} y_k$ [intensity]	$c_{y,k}$	$x_k = m_{x,i} y_k$ [intensity]	$c_{y,k}$	$x_k = m_{x,i} y_k$ [intensity]	$c_{y,k}$
0	0.000000	90	0.255718	177	0.877136
3	0.000000	93	0.269932	180	0.885868
6	0.000000	96	0.287323	183	0.893051
9	0.000000	99	0.309322	186	0.900589
12	0.000000	102	0.331398	189	0.909611
15	0.000000	105	0.351723	192	0.920383
18	0.000000	108	0.369087	195	0.930569
21	0.000000	111	0.384598	198	0.940460
24	0.000000	114	0.400475	201	0.950504
27	0.000103	117	0.416431	204	0.961990
30	0.000809	120	0.434509	207	0.973686
33	0.002872	123	0.455456	210	0.984375
36	0.007526	126	0.480167	213	0.990974
39	0.016357	129	0.507698	216	0.995003
42	0.030720	132	0.533859	219	0.997517
45	0.050907	135	0.556301	222	0.999004
48	0.074230	138	0.578690	225	0.999706
51	0.097958	141	0.604282	228	0.999897
54	0.118965	144	0.631863	231	0.999977
57	0.136944	147	0.658413	234	0.999981
60	0.151184	150	0.684135	237	0.999985
63	0.162281	153	0.712013	240	0.999992
66	0.172062	156	0.743027	243	1.000000
69	0.181019	159	0.770569	246	1.000000
72	0.190571	162	0.792217	249	1.000000
75	0.200386	165	0.809414	252	1.000000
78	0.210571	168	0.824051		
81	0.221203	171	0.837688		
84	0.231625	174	0.852314		
87	0.243053	177	0.865932		

APPENDIX A.5

Characterizing PDFs as the solution to an advective-dispersive process is primarily based on the transformation from the standard-score space to the measurement space. The goal of this transformation is to deconstruct the measurement space distribution into components that characterize the location, scale, and shape of the distribution, while allowing each to evolve independently. From Equation 2.6, the premise of this interpretation is conservation of probability between the measurement space and standard-score space, which can be expressed using the following relationship:

$$\int_{x_0}^{x_1} p_x^* \, dx = \int_{z_0}^{z_1} p_z \, dz$$

From this premise, the derivation of the advective-dispersive process proceeds from definition of the zero-centered PDF $p_x^* \equiv p_x - m_x$. Upon substitution the conservation of probability relationship becomes:

$$\int_{x_0}^{x_1} [p_x - m_x] \, dx = \int_{z_0}^{z_1} p_z \, dz$$

Table 2.1 introduces the relationship between the standard-score variable z and the measurement space variable x. Using a change of variable:

$$z = \frac{m_x - x}{\sigma_x}, \qquad dz = \frac{1}{\sigma_x} dx$$

The conservation of probability expression can be evaluated as:

$$\int_{x_0}^{x_1} [p_x - m_x] \, dx = \int_{x_0}^{x_1} \frac{1}{\sigma_x} p_z \, dx \implies p_x - m_x = \frac{1}{\sigma_x} p_z$$

This results in the probabilistic interpretation of the solution to an advective-dispersive process as:

$$p_x = m_x + \frac{1}{\sigma_x} p_z$$

APPENDIX B.1

Table B.1.1: Residential water consumption data statistics from the histogram data. Note that "bp" denotes billing period and "$acct$" denotes "per account". Water is metered in increments of 1 m^3.

		Period $[t]$	Water Consumption V_t $[m^3/bp]$	Active Accounts N_t $[acct]$	Median Consumption $m_{x,i,t}$ $[m^3/bp/acct]$	Standard Deviation $\sigma_{x,i,t}$ $[m^3/bp/acct]$	Mean Consumption $\mu_{x,i,t}$ $[m^3/bp/acct]$
2007	**Jan/Feb**	1	742,553	20,624	33.00	18.74	36.00
	Mar/Apr	2	700,344	20,681	31.00	18.05	33.86
	May/Jun	3	732,716	20,619	32.00	19.08	35.54
	July/Aug	4	950,893	21,347	40.00	25.65	44.54
	Sept/Oct	5	872,046	21,818	36.00	23.21	39.97
	Nov/Dec	6	781,801	21,497	33.00	19.62	36.37
2008	**Jan/Feb**	7	812,363	22,463	33.00	19.19	36.16
	Mar/Apr	8	812,603	22,526	32.00	20.56	36.07
	May/Jun	9	754,673	21,845	32.00	18.72	34.55
	July/Aug	10	872,729	22,861	34.00	21.90	38.18
	Sept/Oct	11	850,928	23,970	32.00	20.53	35.50
	Nov/Dec	12	802,580	23,408	31.00	18.53	34.29
2009	**Jan/Feb**	13	825,870	24,660	31.00	17.88	33.49
	Mar/Apr	14	832,962	24,687	31.00	18.75	33.74
	May/Jun	15	867,424	25,174	31.00	19.18	34.46
	July/Aug	16	874,294	24,264	33.00	20.45	36.03
	Sept/Oct	17	901,527	26,349	31.00	19.71	34.21
	Nov/Dec	18	897,834	26,799	31.00	18.28	33.50
2010	**Jan/Feb**	19	910,106	26,827	31.00	18.44	33.93
	Mar/Apr	20	913,710	26,650	31.00	19.89	34.29
	May/Jun	21	830,182	25,177	30.00	18.08	32.97
	July/Aug	22	972,815	26,531	33.00	21.93	36.67
	Sept/Oct	23	843,198	25,138	30.00	19.24	33.54
	Nov/Dec	24	907,576	26,882	30.00	19.43	33.76
2011	**Jan/Feb**	25	766,023	23,658	30.00	17.40	32.38
	Mar/Apr	26	867,095	26,830	29.00	18.23	32.32
	May/Jun	27	922,912	26,722	30.00	21.04	34.54
	July/Aug	28	943,924	26,706	32.00	20.72	35.35
	Sept/Oct	29	808,126	23,461	31.00	20.35	34.45
	Nov/Dec	30	877,882	26,848	29.00	18.99	32.70

Table B.1.1: continued

		Period $[t]$	Water Consumption V_t $[m^3/bp]$	Active Accounts N_t $[acct]$	Median Consumption $m_{x,i,t}$ $[m^3/bp/acct]$	Standard Deviation $\sigma_{x,i,t}$ $[m^3/bp/acct]$	Mean Consumption $\mu_{x,i,t}$ $[m^3/bp/acct]$
2012	**Jan/Feb**	31	629,140	20,979	27.00	16.39	29.99
	Mar/Apr	32	833,927	26,866	28.00	17.50	31.04
	May/Jun	33	831,594	26,941	28.00	17.08	30.87
	July/Aug	34	987,313	26,648	33.00	22.47	37.05
	Sept/Oct	35	940,152	26,697	31.00	21.63	35.22
	Nov/Dec	36	839,011	27,014	28.00	17.56	31.06
2013	**Jan/Feb**	37	806,427	27,089	27.00	16.38	29.77
	Mar/Apr	38	814,578	26,960	28.00	16.79	30.21
	May/Jun	39	801,538	26,990	27.00	16.51	29.70
	July/Aug	40	861,573	26,891	29.00	18.83	32.04
	Sept/Oct	41	846,896	26,932	28.00	18.83	31.45
	Nov/Dec	42	836,494	27,225	28.00	17.27	30.73
2014	**Jan/Feb**	43	822,372	27,261	28.00	16.51	30.17
	Mar/Apr	44	787,439	27,049	27.00	16.14	29.11
	May/Jun	45	785,223	26,982	27.00	16.35	29.10
	July/Aug	46	845,542	26,907	28.00	18.57	31.42
	Sept/Oct	47	805,020	27,078	27.00	17.43	29.73
	Nov/Dec	48	810,132	27,268	27.00	16.76	29.71
2015	**Jan/Feb**	49	762,790	27,312	26.00	15.51	27.93
	Mar/Apr	50	782,828	27,142	27.00	16.09	28.84
	May/Jun	51	836,232	27,099	28.00	18.24	30.86
	July/Aug	52	786,989	27,179	27.00	16.24	28.96
	Sept/Oct	53	795,996	27,327	27.00	16.45	29.13
	Nov/Dec	54	814,381	27,145	27.00	17.95	30.00
2016	**Jan/Feb**	55	755,971	27,423	25.00	15.16	27.57
	Mar/Apr	56	742,793	27,297	25.00	14.91	27.21
	May/Jun	57	893,190	27,164	30.00	19.70	32.88
	July/Aug	58	765,496	27,389	26.00	15.34	27.95
	Sept/Oct	59	783,821	27,549	26.00	16.06	28.45
	Nov/Dec	60	831,540	27,207	27.00	18.48	30.56

APPENDIX B.2

Table B.2.1: Optimality fit control function parameters defining the shape of $p_{z,t}$.

		Period $[t]$	$\alpha_{0,t}$ $[-]$	$\alpha_{1,t}$ $[-]$	$\alpha_{2,t}$ $[°]$	$\alpha_{3,t}$ $[-]$	$\alpha_{4,t}$ $[-]$
2007	**Jan/Feb**	1	-0.7691	0.3037	58.27	-0.7429	0.0906
	Mar/Apr	2	-0.7729	0.2801	58.31	-0.6624	0.0639
	May/Jun	3	-0.7429	0.3263	58.38	-0.9118	0.1480
	July/Aug	4	-0.7343	0.3790	60.19	-1.0030	0.1642
	Sept/Oct	5	-0.7325	0.4879	63.15	-1.4010	0.2731
	Nov/Dec	6	-0.7681	0.3296	59.85	-0.8289	0.1045
2008	**Jan/Feb**	7	-0.7589	0.3161	59.07	-0.7835	0.0956
	Mar/Apr	8	-0.6935	0.3652	63.39	-1.1821	0.1839
	May/Jun	9	-0.7789	0.3749	57.84	-0.8504	0.1298
	July/Aug	10	-0.7304	0.3781	61.29	-1.0632	0.1685
	Sept/Oct	11	-0.7095	0.4200	61.86	-1.1166	0.1796
	Nov/Dec	12	-0.7538	0.2922	58.69	-0.7270	0.0822
2009	**Jan/Feb**	13	-0.7918	0.3457	58.53	-0.7430	0.0849
	Mar/Apr	14	-0.7468	0.4009	61.13	-0.9321	0.1201
	May/Jun	15	-0.7402	0.3737	61.78	-1.0471	0.1608
	July/Aug	16	-0.7442	0.4238	61.46	-1.0281	0.1510
	Sept/Oct	17	-0.7555	0.4824	62.98	-1.2716	0.2249
	Nov/Dec	18	-0.7879	0.3699	58.89	-0.7850	0.0941
2010	**Jan/Feb**	19	-0.7824	0.3279	58.88	-0.7384	0.0791
	Mar/Apr	20	0.0622	0.5038	63.41	-1.3298	0.2352
	May/Jun	21	-0.7622	0.3250	58.12	-0.8130	0.1192
	July/Aug	22	-0.7290	0.4853	62.26	-1.2204	0.2120
	Sept/Oct	23	-0.7325	0.4401	62.92	-1.3221	0.2566
	Nov/Dec	24	-0.7246	0.4621	63.76	-1.4259	0.2771
2011	**Jan/Feb**	25	-0.7808	0.3852	59.21	-0.8688	0.1194
	Mar/Apr	26	-0.7388	0.3478	62.28	-0.9823	0.1280
	May/Jun	27	-0.6972	0.5114	65.44	-1.6985	0.3541
	July/Aug	28	-0.7575	0.4712	61.90	-1.2789	0.2451
	Sept/Oct	29	-0.7528	0.5008	63.35	-1.3469	0.2463
	Nov/Dec	30	-0.7260	0.4486	63.12	-1.3182	0.2410

Table B.2.1: continued

		Period $[t]$	$\alpha_{0,t}$ $[-]$	$\alpha_{1,t}$ $[-]$	$\alpha_{2,t}$ $[°]$	$\alpha_{3,t}$ $[-]$	$\alpha_{4,t}$ $[-]$
2012	**Jan/Feb**	31	-0.7602	0.3007	58.72	-0.7975	0.1088
	Mar/Apr	32	-0.7485	0.3618	61.64	-1.0721	0.1724
	May/Jun	33	-0.7325	0.3008	58.39	-0.7794	0.1023
	July/Aug	34	-0.7289	0.4689	63.21	-1.2877	0.2290
	Sept/Oct	35	-0.7466	0.5146	63.98	-1.4837	0.2976
	Nov/Dec	36	-0.7391	0.3717	62.32	-1.1103	0.1774
2013	**Jan/Feb**	37	-0.7699	0.3458	57.97	-0.8304	0.1265
	Mar/Apr	38	-0.7700	0.3462	56.58	-0.7408	0.1029
	May/Jun	39	-0.7521	0.3140	57.64	-0.7209	0.0878
	July/Aug	40	-0.7614	0.3780	58.88	-0.8129	0.1021
	Sept/Oct	41	-0.7226	0.4621	64.67	-1.4570	0.2738
	Nov/Dec	42	-0.7315	0.3938	60.85	-1.0533	0.1735
2014	**Jan/Feb**	43	-0.7887	0.3716	57.05	-0.8385	0.1433
	Mar/Apr	44	-0.7798	0.3235	56.91	-0.5864	0.0411
	May/Jun	45	-0.7876	0.3624	56.41	-0.6889	0.0847
	July/Aug	46	-0.7219	0.3831	62.82	-1.1553	0.1890
	Sept/Oct	47	-0.7684	0.4741	61.16	-1.1494	0.2026
	Nov/Dec	48	-0.7585	0.3655	59.10	-0.8569	0.1197
2015	**Jan/Feb**	49	-0.7585	0.3361	55.45	-0.6681	0.0761
	Mar/Apr	50	-0.7896	0.2852	53.59	-0.5976	0.0767
	May/Jun	51	-0.7863	0.2684	53.50	-0.5246	0.0564
	July/Aug	52	-0.7583	0.4591	55.43	-0.8776	0.1467
	Sept/Oct	53	-0.7858	0.3524	53.36	-0.6478	0.0945
	Nov/Dec	54	-0.8001	0.3490	52.68	-0.6421	0.0924
2016	**Jan/Feb**	55	-0.7855	0.3865	53.73	-0.7608	0.1206
	Mar/Apr	56	-0.7917	0.2663	53.07	-0.5467	0.0662
	May/Jun	57	-0.7863	0.2683	53.50	-0.5246	0.0564
	July/Aug	58	-0.7601	0.4455	54.94	-0.8263	0.1342
	Sept/Oct	59	-0.7850	0.3480	53.47	-0.6358	0.0889
	Nov/Dec	60	-0.7505	0.3969	55.90	-0.8226	0.1307

APPENDIX B.3

Table B.3.1: The CDF of the raw data used to estimate the control function parameters for each bimonthly period in 2007.

	2007					
$y_k[-]$	**Jan/Feb**	**Mar/Apr**	**May/Jun**	**Jul/Aug**	**Sep/Oct**	**Nov/Dec**
0.00	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.25	0.0248	0.0245	0.0230	0.0299	0.0294	0.0264
0.50	0.1182	0.1229	0.1121	0.1327	0.1343	0.1218
0.75	0.2837	0.3020	0.2779	0.2955	0.2985	0.2902
1.00	0.4834	0.4824	0.4773	0.4900	0.4954	0.4846
1.25	0.6795	0.6706	0.6511	0.6558	0.6605	0.6753
1.50	0.8026	0.8034	0.7845	0.7763	0.7778	0.7977
1.75	0.8839	0.8837	0.8685	0.8562	0.8580	0.8741
2.00	0.9320	0.9283	0.9204	0.9110	0.9097	0.9238
2.25	0.9620	0.9569	0.9522	0.9431	0.9399	0.9556
2.50	0.9769	0.9740	0.9708	0.9632	0.9599	0.9720
2.75	0.9866	0.9849	0.9828	0.9753	0.9742	0.9824
3.00	0.9923	0.9899	0.9897	0.9840	0.9841	0.9887
3.25	0.9954	0.9944	0.9944	0.9901	0.9901	0.9932
3.50	0.9975	0.9968	0.9967	0.9945	0.9941	0.9962
3.75	0.9993	0.9989	0.9985	0.9975	0.9971	0.9982
4.00	1.0000	1.0000	0.9995	1.0000	1.0000	1.0000

Table B.3.2: The CDF of the raw data used to estimate the control function parameters for each bimonthly period in 2008.

	2008					
$y_k[-]$	**Jan/Feb**	**Mar/Apr**	**May/Jun**	**Jul/Aug**	**Sep/Oct**	**Nov/Dec**
0.00	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.25	0.0256	0.0234	0.0262	0.0321	0.0253	0.0232
0.50	0.1228	0.1159	0.1261	0.1248	0.1277	0.1216
0.75	0.2870	0.2807	0.2972	0.3008	0.2975	0.2998
1.00	0.4839	0.4765	0.4986	0.4788	0.4942	0.4795
1.25	0.6778	0.6489	0.6730	0.6561	0.6624	0.6615
1.50	0.8000	0.7783	0.7999	0.7746	0.7847	0.7946
1.75	0.8801	0.8606	0.8778	0.8602	0.8624	0.8772
2.00	0.9271	0.9111	0.9272	0.9096	0.9107	0.9221
2.25	0.9577	0.9416	0.9566	0.9422	0.9402	0.9530
2.50	0.9746	0.9606	0.9723	0.9618	0.9603	0.9722
2.75	0.9844	0.9733	0.9839	0.9758	0.9743	0.9827
3.00	0.9904	0.9818	0.9905	0.9841	0.9831	0.9887
3.25	0.9948	0.9883	0.9944	0.9905	0.9892	0.9927
3.50	0.9972	0.9925	0.9973	0.9942	0.9937	0.9957
3.75	0.9988	0.9963	0.9989	0.9971	0.9974	0.9984
4.00	1.0000	0.9990	1.0000	0.9993	1.0000	1.0000

Table B.3.3: The CDF of the raw data used to estimate the control function parameters for each bimonthly period in 2009.

	2009					
$y_k[-]$	**Jan/Feb**	**Mar/Apr**	**May/Jun**	**Jul/Aug**	**Sep/Oct**	**Nov/Dec**
0.00	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.25	0.0247	0.0280	0.0284	0.0385	0.0314	0.0278
0.50	0.1271	0.1328	0.1286	0.1393	0.1377	0.1337
0.75	0.3077	0.3134	0.3010	0.3068	0.3166	0.3155
1.00	0.4948	0.4971	0.4809	0.4973	0.4970	0.4985
1.25	0.6822	0.6777	0.6610	0.6801	0.6682	0.6801
1.50	0.8102	0.8047	0.7880	0.7955	0.7921	0.8063
1.75	0.8898	0.8821	0.8710	0.8731	0.8689	0.8867
2.00	0.9325	0.9245	0.9163	0.9178	0.9130	0.9280
2.25	0.9591	0.9524	0.9474	0.9502	0.9438	0.9561
2.50	0.9758	0.9695	0.9659	0.9671	0.9631	0.9728
2.75	0.9853	0.9806	0.9789	0.9795	0.9764	0.9832
3.00	0.9901	0.9865	0.9861	0.9866	0.9835	0.9888
3.25	0.9938	0.9919	0.9918	0.9922	0.9896	0.9937
3.50	0.9965	0.9951	0.9961	0.9957	0.9942	0.9965
3.75	0.9985	0.9982	0.9987	0.9981	0.9978	0.9985
4.00	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table B.3.4: The CDF of the raw data used to estimate the control function parameters for each bimonthly period in 2010.

	2010					
$y_k[-]$	**Jan/Feb**	**Mar/Apr**	**May/Jun**	**Jul/Aug**	**Sep/Oct**	**Nov/Dec**
0.00	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.25	0.0265	0.0306	0.0339	0.0441	0.0319	0.0284
0.50	0.1265	0.1359	0.1228	0.1480	0.1259	0.1206
0.75	0.3046	0.3169	0.2990	0.3144	0.3042	0.3016
1.00	0.4895	0.4966	0.4844	0.4978	0.4826	0.4807
1.25	0.6709	0.6691	0.6652	0.6714	0.6607	0.6592
1.50	0.8008	0.7931	0.7848	0.7806	0.7737	0.7738
1.75	0.8832	0.8699	0.8731	0.8534	0.8603	0.8600
2.00	0.9246	0.9107	0.9217	0.9008	0.9083	0.9063
2.25	0.9532	0.9405	0.9550	0.9370	0.9428	0.9393
2.50	0.9711	0.9607	0.9717	0.9568	0.9624	0.9595
2.75	0.9827	0.9739	0.9838	0.9716	0.9758	0.9741
3.00	0.9888	0.9829	0.9888	0.9813	0.9846	0.9829
3.25	0.9928	0.9896	0.9937	0.9894	0.9907	0.9900
3.50	0.9962	0.9948	0.9965	0.9948	0.9945	0.9943
3.75	0.9984	0.9977	0.9987	0.9979	0.9982	0.9975
4.00	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table B.3.5: The CDF of the raw data used to estimate the control function parameters for each bimonthly period in 2011.

	2011					
$y_k[-]$	**Jan/Feb**	**Mar/Apr**	**May/Jun**	**Jul/Aug**	**Sep/Oct**	**Nov/Dec**
0.00	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.25	0.0298	0.0345	0.0345	0.0358	0.0323	0.0330
0.50	0.1175	0.1310	0.1269	0.1390	0.1440	0.1335
0.75	0.3030	0.2917	0.3038	0.3055	0.3240	0.2960
1.00	0.4948	0.4787	0.4831	0.4972	0.4976	0.4823
1.25	0.6876	0.6702	0.6501	0.6594	0.6646	0.6620
1.50	0.8015	0.7854	0.7584	0.7753	0.7822	0.7775
1.75	0.8875	0.8648	0.8406	0.8567	0.8615	0.8542
2.00	0.9292	0.9142	0.8877	0.9100	0.9049	0.9023
2.25	0.9583	0.9461	0.9235	0.9415	0.9395	0.9385
2.50	0.9739	0.9648	0.9453	0.9615	0.9595	0.9590
2.75	0.9847	0.9773	0.9644	0.9745	0.9742	0.9724
3.00	0.9904	0.9854	0.9763	0.9832	0.9823	0.9825
3.25	0.9943	0.9919	0.9851	0.9896	0.9889	0.9895
3.50	0.9967	0.9953	0.9916	0.9944	0.9930	0.9938
3.75	0.9991	0.9984	0.9967	0.9978	0.9970	0.9975
4.00	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table B.3.6: The CDF of the raw data used to estimate the control function parameters for each bimonthly period in 2012.

	2012					
$y_k[-]$	**Jan/Feb**	**Mar/Apr**	**May/Jun**	**Jul/Aug**	**Sep/Oct**	**Nov/Dec**
0.00	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.25	0.0278	0.0308	0.0278	0.0440	0.0353	0.0264
0.50	0.1249	0.1283	0.1221	0.1532	0.1535	0.1240
0.75	0.2945	0.2863	0.2875	0.3152	0.3257	0.2894
1.00	0.4734	0.4825	0.4820	0.4916	0.4863	0.4840
1.25	0.6584	0.6538	0.6567	0.6607	0.6499	0.6560
1.50	0.7898	0.7791	0.7849	0.7701	0.7656	0.7785
1.75	0.8713	0.8607	0.8674	0.8468	0.8441	0.8631
2.00	0.9194	0.9139	0.9186	0.8973	0.8909	0.9158
2.25	0.9517	0.9468	0.9498	0.9345	0.9264	0.9467
2.50	0.9700	0.9650	0.9698	0.9552	0.9506	0.9649
2.75	0.9816	0.9777	0.9803	0.9700	0.9673	0.9775
3.00	0.9882	0.9862	0.9878	0.9800	0.9770	0.9852
3.25	0.9927	0.9921	0.9930	0.9874	0.9863	0.9904
3.50	0.9962	0.9956	0.9959	0.9928	0.9920	0.9946
3.75	0.9984	0.9983	0.9984	0.9972	0.9964	0.9974
4.00	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table B.3.7: The CDF of the raw data used to estimate the control function parameters for each bimonthly period in 2013.

	2013					
$y_k[-]$	**Jan/Feb**	**Mar/Apr**	**May/Jun**	**Jul/Aug**	**Sep/Oct**	**Nov/Dec**
0.00	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.25	0.0286	0.0312	0.0314	0.0413	0.0328	0.0276
0.50	0.1297	0.1318	0.1348	0.1469	0.1371	0.1251
0.75	0.3053	0.3023	0.3067	0.3121	0.3021	0.2950
1.00	0.4829	0.4991	0.4816	0.4949	0.4875	0.4912
1.25	0.6618	0.6713	0.6627	0.6731	0.6490	0.6639
1.50	0.7921	0.7951	0.7920	0.7832	0.7668	0.7878
1.75	0.8750	0.8727	0.8773	0.8576	0.8500	0.8663
2.00	0.9198	0.9242	0.9195	0.9089	0.8996	0.9172
2.25	0.9507	0.9555	0.9519	0.9435	0.9333	0.9478
2.50	0.9702	0.9719	0.9698	0.9626	0.9555	0.9670
2.75	0.9820	0.9824	0.9814	0.9745	0.9709	0.9789
3.00	0.9884	0.9893	0.9876	0.9832	0.9799	0.9865
3.25	0.9931	0.9935	0.9923	0.9900	0.9867	0.9923
3.50	0.9964	0.9964	0.9961	0.9948	0.9923	0.9958
3.75	0.9986	0.9986	0.9983	0.9975	0.9965	0.9984
4.00	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table B.3.8: The CDF of the raw data used to estimate the control function parameters for each bimonthly period in 2014.

	2014					
$y_k[-]$	**Jan/Feb**	**Mar/Apr**	**May/Jun**	**Jul/Aug**	**Sep/Oct**	**Nov/Dec**
0.00	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.25	0.0291	0.0344	0.0370	0.0346	0.0353	0.0292
0.50	0.1279	0.1408	0.1479	0.1377	0.1475	0.1369
0.75	0.2986	0.3211	0.3244	0.2958	0.3281	0.3150
1.00	0.4992	0.4980	0.4977	0.4807	0.4971	0.4880
1.25	0.6697	0.6760	0.6737	0.6467	0.6647	0.6654
1.50	0.7975	0.8049	0.8010	0.7679	0.7884	0.7903
1.75	0.8782	0.8847	0.8827	0.8500	0.8664	0.8736
2.00	0.9256	0.9247	0.9249	0.9012	0.9105	0.9166
2.25	0.9559	0.9567	0.9533	0.9357	0.9430	0.9485
2.50	0.9741	0.9737	0.9719	0.9582	0.9628	0.9672
2.75	0.9835	0.9824	0.9824	0.9719	0.9751	0.9791
3.00	0.9899	0.9880	0.9882	0.9818	0.9827	0.9874
3.25	0.9940	0.9930	0.9931	0.9887	0.9898	0.9919
3.50	0.9971	0.9967	0.9963	0.9935	0.9944	0.9953
3.75	0.9988	0.9987	0.9983	0.9973	0.9974	0.9980
4.00	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table B.3.9: The CDF of the raw data used to estimate the control function parameters for each bimonthly period in 2015.

	2015					
$y_k[-]$	**Jan/Feb**	**Mar/Apr**	**May/Jun**	**Jul/Aug**	**Sep/Oct**	**Nov/Dec**
0.00	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.25	0.0345	0.0371	0.0373	0.0354	0.0336	0.0383
0.50	0.1306	0.1467	0.1410	0.1481	0.1445	0.1545
0.75	0.3164	0.3302	0.3084	0.3289	0.3253	0.3279
1.00	0.5010	0.5037	0.4925	0.5009	0.5036	0.4915
1.25	0.6850	0.6795	0.6578	0.6805	0.6772	0.6586
1.50	0.7991	0.8057	0.7760	0.8045	0.8022	0.7785
1.75	0.8818	0.8859	0.8572	0.8838	0.8793	0.8591
2.00	0.9241	0.9282	0.9080	0.9246	0.9218	0.9019
2.25	0.9564	0.9560	0.9412	0.9545	0.9513	0.9360
2.50	0.9722	0.9741	0.9609	0.9730	0.9702	0.9573
2.75	0.9831	0.9842	0.9744	0.9829	0.9814	0.9724
3.00	0.9885	0.9888	0.9831	0.9881	0.9874	0.9819
3.25	0.9934	0.9934	0.9887	0.9931	0.9920	0.9890
3.50	0.9963	0.9958	0.9935	0.9961	0.9959	0.9936
3.75	0.9982	0.9984	0.9973	0.9983	0.9983	0.9976
4.00	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table B.3.10: The CDF of the raw data used to estimate the control function parameters for each bimonthly period in 2016.

	2016					
$y_k[-]$	**Jan/Feb**	**Mar/Apr**	**May/Jun**	**Jul/Aug**	**Sep/Oct**	**Nov/Dec**
0.00	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.25	0.0338	0.0400	0.0430	0.0350	0.0322	0.0348
0.50	0.1330	0.1374	0.1481	0.1317	0.1285	0.1477
0.75	0.2929	0.2971	0.3301	0.3157	0.3133	0.3242
1.00	0.4780	0.4814	0.5043	0.4959	0.4931	0.4866
1.25	0.6706	0.6795	0.6746	0.6827	0.6717	0.6478
1.50	0.7892	0.7971	0.7776	0.7965	0.7863	0.7660
1.75	0.8691	0.8761	0.8597	0.8818	0.8718	0.8478
2.00	0.9178	0.9242	0.9075	0.9270	0.9182	0.8946
2.25	0.9534	0.9569	0.9405	0.9580	0.9483	0.9295
2.50	0.9714	0.9742	0.9594	0.9735	0.9670	0.9529
2.75	0.9815	0.9845	0.9740	0.9842	0.9790	0.9683
3.00	0.9892	0.9897	0.9830	0.9901	0.9868	0.9786
3.25	0.9938	0.9940	0.9899	0.9940	0.9923	0.9868
3.50	0.9964	0.9962	0.9940	0.9969	0.9949	0.9928
3.75	0.9983	0.9983	0.9968	0.9987	0.9975	0.9967
4.00	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

APPENDIX B.4

Table B.4.1: Real water price and weather data. Price is discounted to 2004$ using CPI. Temperature is average daily temperature in each billing period. Rainfall is the number of days within the billing period with less than 2 mm precipitation.

		Period $[t]$	Real Water Price P_t $[2004-\$]$	Average Daily High Temperature T_t [℃]	Days < 2mm Precipitation R_t $[days]$	Weather Score W_t $[days \cdot ℃]$
2007	Jan/Feb	1	1.91	-3.35	54	-180.90
	Mar/Apr	2	1.91	6.45	46	296.70
	May/Jun	3	1.91	23.05	44	1037.25
	July/Aug	4	1.91	25.35	45	1140.75
	Sept/Oct	5	1.91	20.55	52	1068.60
	Nov/Dec	6	1.91	2.20	56	121.00
2008	Jan/Feb	7	2.12	-1.35	55	-74.25
	Mar/Apr	8	2.12	7.40	52	384.80
	May/Jun	9	2.12	19.65	41	825.30
	July/Aug	10	2.12	24.90	48	1195.20
	Sept/Oct	11	2.12	17.55	46	807.30
	Nov/Dec	12	2.12	2.30	54	121.90
2009	Jan/Feb	13	2.43	-2.95	51	-150.45
	Mar/Apr	14	2.43	8.55	43	367.65
	May/Jun	15	2.43	20.10	41	844.20
	July/Aug	16	2.43	23.45	45	1031.80
	Sept/Oct	17	2.43	16.35	43	719.40
	Nov/Dec	18	2.43	4.30	46	193.50
2010	Jan/Feb	19	2.57	-2.65	28	-121.90
	Mar/Apr	20	2.57	11.80	46	542.80
	May/Jun	21	2.57	20.55	37	801.45
	July/Aug	22	2.57	27.10	39	1056.90
	Sept/Oct	23	2.57	17.20	42	722.40
	Nov/Dec	24	2.57	2.85	52	145.35
2011	Jan/Feb	25	2.69	-3.45	46	-169.05
	Mar/Apr	26	2.69	6.65	50	259.35
	May/Jun	27	2.69	20.70	38	890.10
	July/Aug	28	2.69	27.60	45	1269.60
	Sept/Oct	29	2.69	17.45	48	645.65
	Nov/Dec	30	2.69	6.25	48	293.75

Table B.4.1: continued

		Period $[t]$	Real Water Price P_t $[2004 - \$]$	Average Daily High Temperature T_t [℃]	Days < 2mm Precipitation R_t $[days]$	Weather Score W_t $[days \cdot ℃]$
2012	Jan/Feb	31	2.81	0.80	51	37.60
	Mar/Apr	32	2.81	12.00	50	600.00
	May/Jun	33	2.81	23.65	39	1135.20
	July/Aug	34	2.81	27.55	49	1267.30
	Sept/Oct	35	2.81	17.05	51	613.80
	Nov/Dec	36	2.81	4.35	51	208.80
2013	Jan/Feb	37	2.95	-1.20	48	-55.20
	Mar/Apr	38	2.95	6.00	47	288.00
	May/Jun	39	2.95	21.45	43	879.45
	July/Aug	40	2.95	25.55	44	1251.95
	Sept/Oct	41	2.95	17.95	52	771.85
	Nov/Dec	42	2.95	1.25	55	62.50
2014	Jan/Feb	43	3.03	-5.95	44	-291.55
	Mar/Apr	44	3.03	4.80	44	206.40
	May/Jun	45	3.03	22.00	48	946.00
	July/Aug	46	3.03	24.15	51	1062.60
	Sept/Oct	47	3.03	17.20	43	705.20
	Nov/Dec	48	3.03	2.65	51	140.45
2015	Jan/Feb	49	3.11	-2.51	51	-128.14
	Mar/Apr	50	3.11	7.96	45	358.03
	May/Jun	51	3.11	21.39	47	1005.51
	July/Aug	52	3.11	25.71	55	1413.84
	Sept/Oct	53	3.11	17.66	47	830.14
	Nov/Dec	54	3.11	3.27	46	150.36
2016	Jan/Feb	55	3.15	-0.05	50	-2.50
	Mar/Apr	56	3.15	8.25	41	338.25
	May/Jun	57	3.15	22.55	47	1059.85
	July/Aug	58	3.15	28.20	45	1269.00
	Sept/Oct	59	3.15	20.10	45	904.50
	Nov/Dec	60	3.15	5.25	31	162.75

APPENDIX B.5

Table B.5.1: Transport model results.

			Transport Model Results		
		Period $[t]$	m_x $[m^3/bp/acct]$	σ_x $[m^3/bp/acct]$	$\hat{\mu}_{x\|\hat{p}_x} = m_x + \sigma_x \mu_z$ $[m^3/bp/acct]$
2007	**Jan/Feb**	1	34.3394	19.4993	37.1244
	Mar/Apr	2	32.3931	19.7397	35.2947
	May/Jun	3	35.6966	25.2660	39.2248
	July/Aug	4	36.7705	26.5374	40.4112
	Sept/Oct	5	36.0060	25.6381	39.5672
	Nov/Dec	6	32.7374	19.3482	35.5041
2008	**Jan/Feb**	7	32.3874	18.7059	35.1844
	Mar/Apr	8	31.4284	19.2316	34.4148
	May/Jun	9	32.7940	21.8090	36.0077
	July/Aug	10	35.6701	25.5413	39.2025
	Sept/Oct	11	32.6943	21.6639	35.8955
	Nov/Dec	12	31.6802	18.6608	34.4809
2009	**Jan/Feb**	13	30.8203	17.9453	33.5267
	Mar/Apr	14	30.0188	17.9623	32.8856
	May/Jun	15	31.1551	20.1228	34.2067
	July/Aug	16	32.0949	21.5371	35.2591
	Sept/Oct	17	30.6840	19.3584	33.6732
	Nov/Dec	18	30.0514	17.6856	32.8870
2010	**Jan/Feb**	19	29.8910	17.4934	32.5529
	Mar/Apr	20	29.6036	17.9055	32.4644
	May/Jun	21	30.2468	19.0731	33.2027
	July/Aug	22	31.2973	20.7801	34.3831
	Sept/Oct	23	30.0053	18.6564	32.9283
	Nov/Dec	24	29.4396	17.2111	32.2256
2011	**Jan/Feb**	25	29.2548	17.2692	32.0037
	Mar/Apr	26	28.8360	16.8502	31.5939
	May/Jun	27	29.8688	18.8709	32.7955
	July/Aug	28	31.4368	21.6003	34.5545
	Sept/Oct	29	29.2354	17.7149	32.0747
	Nov/Dec	30	28.8416	16.8794	31.6040

Table B.5.1: continued

		Period $[t]$	Transport Model Results		
			m_x $[m^3/bp/acct]$	σ_x $[m^3/bp/acct]$	$\hat{\mu}_{x\|\hat{p}_x} = m_x + \sigma_x \mu_z$ $[m^3/bp/acct]$
2012	**Jan/Feb**	31	28.2909	28.2909	30.9941
	Mar/Apr	32	28.5713	28.5713	31.3430
	May/Jun	33	29.8681	29.8681	32.8170
	July/Aug	34	30.3427	30.3427	33.3522
	Sept/Oct	35	28.5921	28.5921	31.3670
	Nov/Dec	36	28.2588	28.2588	30.9701
2013	**Jan/Feb**	37	27.5722	27.5722	30.2321
	Mar/Apr	38	27.6573	27.6573	30.3288
	May/Jun	39	28.3352	28.3352	31.1064
	July/Aug	40	29.1071	29.1071	31.9930
	Sept/Oct	41	28.1618	28.1618	30.9074
	Nov/Dec	42	27.5759	27.5759	30.2359
2014	**Jan/Feb**	43	27.0660	27.0660	29.7102
	Mar/Apr	44	27.2432	27.2432	29.8825
	May/Jun	45	27.9643	27.9643	30.7035
	July/Aug	46	28.1279	28.1279	30.8946
	Sept/Oct	47	27.6694	27.6694	30.3620
	Nov/Dec	48	27.2055	27.2055	29.8420
2015	**Jan/Feb**	49	26.6081	26.6081	29.2184
	Mar/Apr	50	26.9888	26.9888	29.6092
	May/Jun	51	27.5231	27.5231	30.2232
	July/Aug	52	27.8761	27.8761	30.8379
	Sept/Oct	53	27.3753	27.3753	30.0463
	Nov/Dec	54	26.8240	26.8240	29.4349
2016	**Jan/Feb**	55	26.4878	26.4878	29.0839
	Mar/Apr	56	26.7936	26.7936	29.3984
	May/Jun	57	27.2921	27.2921	29.9744
	July/Aug	58	27.3988	27.3988	30.1184
	Sept/Oct	59	27.2019	27.2019	29.8607
	Nov/Dec	60	26.6417	26.6417	29.2399

Table B.5.2: Direct regression model results.

		Period $[t]$	$\widehat{\mu}_x$ $[m^3/bp/acct]$
2007	**Jan/Feb**	1	36.1357
	Mar/Apr	2	36.3357
	May/Jun	3	39.9083
	July/Aug	4	40.7235
	Sept/Oct	5	40.1470
	Nov/Dec	6	36.0703
2008	**Jan/Feb**	7	34.8673
	Mar/Apr	8	35.3078
	May/Jun	9	36.9553
	July/Aug	10	39.2652
	Sept/Oct	11	36.8645
	Nov/Dec	12	34.8961
2009	**Jan/Feb**	13	33.1917
	Mar/Apr	14	33.4527
	May/Jun	15	34.7924
	July/Aug	16	35.6087
	Sept/Oct	17	34.3397
	Nov/Dec	18	33.2260
2010	**Jan/Feb**	19	32.4234
	Mar/Apr	20	32.9784
	May/Jun	21	33.6682
	July/Aug	22	34.6100
	Sept/Oct	23	33.4292
	Nov/Dec	24	32.4358
2011	**Jan/Feb**	25	31.7897
	Mar/Apr	26	31.8550
	May/Jun	27	33.0802
	July/Aug	28	34.4652
	Sept/Oct	29	32.4458
	Nov/Dec	30	31.8872
2012	**Jan/Feb**	31	31.0688
	Mar/Apr	32	31.5658
	May/Jun	33	32.8528
	July/Aug	34	33.2926
	Sept/Oct	35	31.5890
	Nov/Dec	36	31.1273
2013	**Jan/Feb**	37	30.3236
	Mar/Apr	38	30.4074
	May/Jun	39	31.1321
	July/Aug	40	31.9654
	Sept/Oct	41	30.9456
	Nov/Dec	42	30.3245
2014	**Jan/Feb**	43	29.9588
	Mar/Apr	44	29.9226
	May/Jun	45	30.6504
	July/Aug	46	30.8504
	Sept/Oct	47	30.3109
	Nov/Dec	48	29.9031
2015	**Jan/Feb**	49	29.4609
	Mar/Apr	50	29.5344
	May/Jun	51	30.1147
	July/Aug	52	30.7642
	Sept/Oct	53	29.9031
	Nov/Dec	54	29.4650
2016	**Jan/Feb**	55	29.2320
	Mar/Apr	56	29.2960
	May/Jun	57	29.8601
	July/Aug	58	30.1324
	Sept/Oct	59	29.6895
	Nov/Dec	60	29.2468

APPENDIX C.1

To begin, define the probabilistic radius $\mathbb{R}$ to be equal to the one-dimensional radial CDF and the probabilistic area $\mathbb{P}$ to be angle θ multiplied by the probabilistic area squared $\mathbb{R}^2$ using a unit circle analogy as:

$$\int \frac{dp_{r,\mathrm{t}}^*}{dn} dr = 1, \qquad \mathbb{R} \equiv \int \frac{dp_{r,\mathrm{t}}^*}{dn} dr, \qquad \mathbb{P} \equiv \theta \mathbb{R}^2$$

A third-order derivative of probabilistic area $d\mathbb{P}$ results from taking the derivative twice with respect to change in radius dr and once with respect to change in angle $d\theta$:

$$\frac{d^2\mathbb{P}}{d\theta dr} = 2\frac{d\mathbb{R}}{dr}\mathbb{R} \implies \frac{d^3\mathbb{P}}{d\theta dr dr} = 2\frac{d}{dr}\left[\frac{d\mathbb{R}}{dr}\right]\mathbb{R} + 2\left[\frac{d\mathbb{R}}{dr}\right]^2$$

After evaluating the derivative of the probabilistic radius $\mathbb{R}$ with respect to change in radius dr, this progression makes a substitution to relate the area-based and radial PDFs.

$$\frac{d\mathbb{R}}{dr} = \frac{dp_{r,\mathrm{t}}^*}{dn}, \qquad \frac{d}{dr}\left[\frac{d\mathbb{R}}{dr}\right] = \frac{d^2p_{r,\mathrm{t}}^*}{dndr} \implies \frac{d^3\mathbb{P}}{d\theta dr dr} = 2\frac{d^2p_{r,\mathrm{t}}^*}{dndr}\mathbb{R} + 2\left[\frac{dp_{r,\mathrm{t}}^*}{dn}\right]^2$$

The only unknown in this relationship is the derivative $\frac{d^2p_{r,\mathrm{t}}^*}{dndr}$, which is analogous to the rate of change of probability density along the radial axis. The progression transforms this relationship into the standard-score space as follow:

$$\frac{dp_{r,\mathrm{t}}^*}{dn} = \frac{1}{\sqrt{s_{r,t,n}^2}}\bar{p}_z, \qquad \frac{d^2p_{r,\mathrm{t}}^*}{dndz} = \frac{1}{\sqrt{s_{r,t,n}^2}}\frac{d\bar{p}_z}{dz}$$

The PDF $\frac{dp_{r,\mathrm{t}}^*}{dn}$ is a scaled normal distribution, which can be evaluated as a derivative with respect to the standard-score space $\frac{d^2p_{r,\mathrm{t}}^*}{dndz}$. The derivative of the normal distribution in the standard-score space is a well-known relationship as:

$$\because \frac{d\bar{p}_z}{dz} = -z\bar{p}_z, \qquad \therefore \frac{d^2p_{r,\mathrm{t}}^*}{dndz} = -\frac{z}{\sqrt{s_{r,t,n}^2}}\bar{p}_z$$

Upon substitution of $\frac{d\bar{p}_z}{dz}$ into the relationship for $\frac{d^2p^*_{r,\mathrm{t}}}{dndz}$, a change of variable allows determination of $\frac{d^2p^*_{r,\mathrm{t}}}{dndr}$:

$$r = m_{r,t,n} + \sqrt{s^2_{r,t,n}}z, \qquad dr = \sqrt{s^2_{r,t,n}}dz$$

$$\because \frac{d^2p^*_{r,\mathrm{t}}}{dndr} = \frac{1}{\sqrt{s^2_{r,t,n}}}\frac{d^2\bar{p}_z}{dndz}, \qquad \therefore \frac{d^2p^*_{r,\mathrm{t}}}{dndr} = -\frac{z}{s^2_{r,t,n}}\bar{p}_z$$

Upon substitution of $\frac{d^2p^*_{r,\mathrm{t}}}{dndr} = -\frac{z}{s^2_{r,t,n}}\bar{p}_z$, the resulting relationship can interpret the probabilistic area $\mathbb{P}$ through integration as:

$$\mathbb{P} = 2\iiint \left\{-\frac{z}{s^2_{r,n,t}}\bar{p}_z\mathbb{R} + \left[\frac{dp^*_{r,\mathrm{t}}}{dn}\right]^2\right\} drdrd\theta$$

$$\mathbb{P} = -\frac{2}{s^2_{r,t,n}}\iiint [z\bar{p}_z\mathbb{R}]\, dzdzd\theta + 2\iiint \left[\frac{dp^*_{r,\mathrm{t}}}{dn}\right]^2 drdrd\theta$$

Here, the progression can evaluate $2s^2_{r,t,n}\iiint z\bar{p}_z\, dzdzd\theta$ to be analogous to the arithmetic mean of the distribution in the standard-score space. Ultimately, when the distribution $\frac{dp^*_{r,\mathrm{t}}}{dn}$ is symmetrical, such as the case when it is a scaled normal distribution, $\int z\bar{p}_z\, dz$ will always be zero resulting in the following equality:

$$\mathbb{R} = 1, \qquad \int z\bar{p}_z\, dz = 0 \implies \mathbb{P} = 2\iiint \left[\frac{dp^*_{r,\mathrm{t}}}{dn}\right]^2 drdrd\theta$$

$$d\mathbb{A} \equiv drdrd\theta \implies \mathbb{P} = 2\int \left[\frac{dp^*_{r,\mathrm{t}}}{dn}\right]^2 d\mathbb{A}$$

Upon substitution, the probabilistic area is twice the area integral of the squared radial PDF, $\mathbb{P} = 2\int \left[\frac{dp^*_{r,\mathrm{t}}}{dn}\right]^2 d\mathbb{A}$.

APPENDIX C.2

To begin, the progression considers a normal distribution scaled by the square root of variance $\sqrt{s_{1,t,n}^2}$ as:

$$p_{1,t}^* = \frac{1}{\sqrt{s_{1,t,n}^2}} \bar{p}_z$$

Upon substitution of the normal distribution:

$$\bar{p}_z = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} z^2\right) \implies p_{1,t}^* = \frac{1}{\sqrt{2\pi s_{1,t,n}^2}} \exp\left(-\frac{1}{2} \frac{\left[r - m_{1,t,n}\right]^2}{s_{1,t,n}^2}\right)$$

Squaring this relationship allows this progression to estimate the resulting PDF from multiplying two normal distributions together:

$$p_{1,t}^* p_{2,t}^* = \frac{1}{2\pi \sqrt{s_{1,t,n}^2 s_{2,t,n}^2}} \exp\left(-\frac{1}{2}\left[r - m_{1,t,n}\right]^2 \frac{1}{s_{1,t,n}^2} - \frac{1}{2}\left[r - m_{2,t,n}\right]^2 \frac{1}{s_{2,t,n}^2}\right)$$

Assuming the two normal distributions have the same center-point simplifies this relationship.

$$m_{1,t,n} = m_{2,t,n} \implies p_{1,t}^* p_{2,t}^* = \frac{1}{2\pi \sqrt{s_{1,t,n}^2 s_{2,t,n}^2}} \exp\left(-\frac{1}{2}\left[r - m_{1,t,n}\right]^2 \left[\frac{s_{1,t,n}^2 + s_{2,t,n}^2}{s_{1,t,n}^2 s_{2,t,n}^2}\right]\right)$$

$$p_{1,t}^* p_{2,t}^* = \frac{1}{2\pi \sqrt{s_{1,t,n}^2 s_{2,t,n}^2}} \exp\left(-\frac{1}{2} \frac{\left[r - m_{1,t,n}\right]^2}{\left[\frac{s_{1,t,n}^2 s_{2,t,n}^2}{s_{1,t,n}^2 + s_{2,t,n}^2}\right]}\right)$$

Using the same argument as Bromiley (2003), this result can be interpreted in the context of a normal distribution as:

$$s_{1,2,t,n}^2 = \frac{s_{1,t,n}^2 s_{2,t,n}^2}{s_{1,t,n}^2 + s_{2,t,n}^2} \implies p_{1,t}^* p_{2,t}^* = \frac{1}{2\pi \sqrt{s_{1,t,n}^2 s_{2,t,n}^2}} \exp\left(-\frac{1}{2} \frac{\left[r - m_{1,t,n}\right]^2}{s_{1,2,t,n}^2}\right)$$

However, there is obviously a disconnect between the variance terms that scale the normal distribution, $\frac{1}{2\pi\sqrt{s_{1,t,n}^2 s_{2,t,n}^2}}$, and the variance term that gives the distribution its shape, $-\frac{1}{2}\frac{[r-m_{1,t,n}]^2}{s_{1,2,t,n}^2}$. The normal distribution in the standard-score space has the property that it is identical for all applications. This relationship ensures there is a translation between all variance terms relating to a normal distribution.

$$For\ m_{1,t,n}=0, \qquad z=\frac{r}{\sqrt{s_{1,2,t,n}^2}}, \qquad z=\frac{r}{\sqrt{s_{1,t,n}^2}} \implies \frac{r}{\sqrt{s_{1,2,t,n}^2}}=\frac{r}{\sqrt{s_{1,t,n}^2}}$$

This transformation is notable because it represents a transformation from two-dimensional normal distribution into an analogous one-dimensional distribution. Effectively, the analysis needs to conserve probability for a two-dimensional random walk. This transformation provides a form that is conducive to Einstein's diffusion PDE. Applying this interpretation and setting the variance terms of the original normal distributions to be equal produces a scaled normal distribution for a polar-coordinate system in two-dimensions:

$$s_{1,t,n}^2=s_{r,t,n}^2, \qquad p_{1,t}^* p_{2,t}^* = \left[\frac{dp_{r,t}^*}{dn}\right]^2 \implies \left[p_{r,t}^*\right]^2=\frac{1}{2\pi s_{r,t,n}^2}\exp\left(-\frac{1}{2}\frac{\left[r-m_{r,t,n}\right]^2}{s_{r,t,n}^2}\right)$$

This exercise is completed by substituting the Einstein-Smoluchowski equation for the variance term and show that this approach produces a solution that is analogous to the Fourier heat equation from Carslaw (1921).

$$s_{1,t,n}^2=2D_{r,t,n}t \implies \left[\frac{dp_{r,t}^*}{dn}\right]^2=\frac{1}{4\pi D_{r,t,n}t}\exp\left(-\frac{\left[r-m_{1,t,n}\right]^2}{4D_{r,t,n}t}\right)$$

Taking the time-integral of this relationship produces the exact solution to the Fourier heat equation.

$$\int\left[\frac{dp_{r,t}^*}{dn}\right]^2 dt=\int\frac{1}{4\pi D_{r,t,n}t}\exp\left(-\frac{\left[r-m_{1,t,n}\right]^2}{4D_{r,t,n}t}\right)dt$$

This is the expected form of this solution to express the physical nature of diffusion.

APPENDIX C.3

This analysis considers the time-derivative of the probabilistic solution to two-dimensional diffusion (Equation 9) and evaluates how the area-based PDF will change with respect to time $\frac{dp_{\mathbb{P},t}}{dt}$ as:

$$\frac{dp_{\mathbb{A},t,n}}{dt} = \underbrace{\frac{dm_{\mathbb{A},t,n}}{dt}}_{\substack{probabilistic \\ advective \\ process}} \pm \underbrace{\frac{d}{dt}\left[\frac{1}{\sigma_{\mathbb{A},t,n}}\right]\bar{p}_z}_{\substack{probabilistic \\ dispersive \\ process}}, \qquad \frac{d\bar{p}_z}{dt} = 0, \qquad \sigma_{\mathbb{P},t,n} = \sqrt{s^2_{\mathbb{A},t,n}}$$

Where, the notation for the standard deviation $\sigma_{\mathbb{A},t,n}$ is consistent with Chapter 2 and characterizes the time-dependent dispersive process. The only notable difference is that the dispersive process is applied to the constant shape of a normal distribution $\bar{p}_z$. Evaluating the time-derivative based on the square-root of variance value from the Einstein-Smoluchowski equation allows this progression to characterize how the two-dimensional diffusion PDF will evolve through time. Notice that the time-derivative of the inverse standard deviation is always non-zero, $\frac{d}{dt}\left[\frac{1}{\sigma_{\mathbb{A},t,n}}\right] \neq 0$, because it is dependent upon the diffusion coefficient from as $\sigma_{\mathbb{A},t,n} = \sqrt{\frac{\pi}{2}} D_{r,t,n} t$, as:

$$\frac{1}{\sigma_{\mathbb{A},t,n}} = \frac{1}{\sqrt{\frac{\pi}{2}} D_{r,t,n}\, t} \implies \frac{d}{dt}\left[\frac{1}{\sigma_{\mathbb{A},t,n}}\right] = \pm \frac{1}{\sqrt{\frac{\pi}{2}} D_{r,t,n} t^2}$$

This relationship characterizes how the probabilistic solution in two dimensions will evolve with respect to some time-interval $\mathcal{T}$ as:

$$p^*_{\mathbb{A},t,n} = p^*_{\mathbb{A},0,n} + \frac{dp^*_{\mathbb{A},t,n}}{dt}\Delta t \implies p_{\mathbb{A},t,n} = m_0 + \int_0^{\mathcal{T}} \frac{dm_{\mathbb{A},t,n}}{dt} dt \pm \left[\int_0^{\mathcal{T}} \frac{1}{\sqrt{\frac{\pi}{2}} D_{r,t,n} t^2} \mathrm{d}t\right]\bar{p}_z$$

$$p_{\mathbb{A},t,n} = m_{\mathbb{A},0,n} + \frac{dm_{\mathbb{A},t,n}}{dt}\mathcal{T} \pm \left[\frac{1}{\sqrt{\frac{\pi}{2}} D_{r,t,n} \mathcal{T}}\right]\bar{p}_z$$

Defining the time-derivative for the advective-dispersive process fully characterizes how this probabilistic solution will evolve with respect to time. Plotting this relationship alongside iterations of a random walk can provide insight into the probabilistic solution to the advective-dispersive process. The random walk numerical examples from Appendix C use this methodology for computing the corresponding probabilistic solutions.

PROBABILISTIC TWO-DIMENSIONAL RANDOM WALK

This section will develop a numerical example of Perrin's suspended solids experiment to provide a discrete interpretation of the solution to the area-based probabilistic PDE. Here, the analysis considers a random walk as a discrete interpretation of the probabilistic solution for a single molecule. Plotting both the random walk and its probabilistic counterpart on the same figure illustrates how the PDE from Equation 9 actually predicts the probability that the random walk will exist within some radius of its origin solely from the influence of diffusion. Notably, a random walk governed by a normal distribution has an expected value of zero; however, the solution to the Fourier equation suggests that the probability the random walk will pass through the origin is continually decreasing. In this spirit, the expected value of the probabilistic advective-dispersive process can be evaluated, denoted with triangle brackets $\langle \quad \rangle$. Naturally, the expected value of a zero-centered normal distribution is constrained to be zero due to area-based symmetry, $\langle \frac{dp^*_{\mathbb{A},t,n}}{dn} \rangle = \langle \frac{1}{\sqrt{\frac{1}{2\pi} D_{r,t,n} t}} \bar{p}_z \rangle \implies \langle \frac{dp^*_{\mathbb{A},t,n}}{dn} \rangle = 0$. Similarly, the expected value of a diffusive process $p_{\mathbb{A},t,n}$ with a nonzero central-tendency is equal to the initial position for a system at rest, $\langle \frac{dp^*_{\mathbb{A},t,n}}{dn} \rangle = m_0$.

This analysis considers two random variable to create a discrete random walk that is consistent with the probabilistic solution: $\Theta \in \mathcal{U}(0, \pi)$ and $\mathcal{L} \in \mathrm{N}\left(0, \sqrt{s^2_{\ell,t,n}}\right)$ – where, Θ is a uniform distribution that represents direction and $\mathcal{L}$ is a normal distribution that represents the displacement of the random walk. Notably, it is assumed that each displacement occurs over some constant unit of time, such that the displacement variable $\mathcal{L}$ represents of the relative speed of displacement. Therefore, the combination of displacement speed $\mathcal{L}$ and direction Θ implies velocity. Although the random variables represent polar coordinates, this relationship projects onto a Cartesian plane for visualization.

$$\frac{da}{dt} = \frac{dm_{\mathbb{A},t,n}}{dt}\cos(\theta) + \frac{d\mathcal{L}}{dt}\cos(\Theta)\,, \qquad \frac{db}{dt} = \frac{dm_{\mathbb{A},t,n}}{dt}\sin(\theta) + \frac{d\mathcal{L}}{dt}\sin(\Theta),$$

$$x_0 = 0, \qquad y_0 = 0, \qquad coor(a,b) = coor\left(\, a_0 + \frac{da}{dt}\Delta t, b_0 + \frac{db}{dt}\Delta t\right)$$

Where, $\frac{da}{dt}$ and $\frac{db}{dt}$ represent the velocity of travel in the a- and b-directions, respectively (x, y, and z variables for the probabilistic interpretation); $\frac{dm_{\mathbb{A},t,n}}{dt}$ represents a predefined deterministic step size in direction θ. This numerical example of a random walk applies 250 time-steps from origin location $coor(0,0)$, while applying the uniform and Gaussian distributions that describe velocity. Notably, this example considers no deterministic contribution to the random walk, $\frac{dm_{\mathbb{A},t,n}}{dt} = 0$. The numerical example presented in Figure C.3.1 reflects a standard normal random walk with equal chance of moving in a positive or negative direction $0 \leq \theta \leq \pi$. This example provides sufficient information to evaluate an Einsteinian diffusion coefficient $D_{\ell,t,n}$ and the resulting radial diffusion coefficient $D_{r,t,n}$. Upon evaluating the radial diffusion coefficient, this exercise can parameterize the dispersive process and visualize the resulting probabilistic area at time $\mathcal{T}$ on the same figure as the random walk.

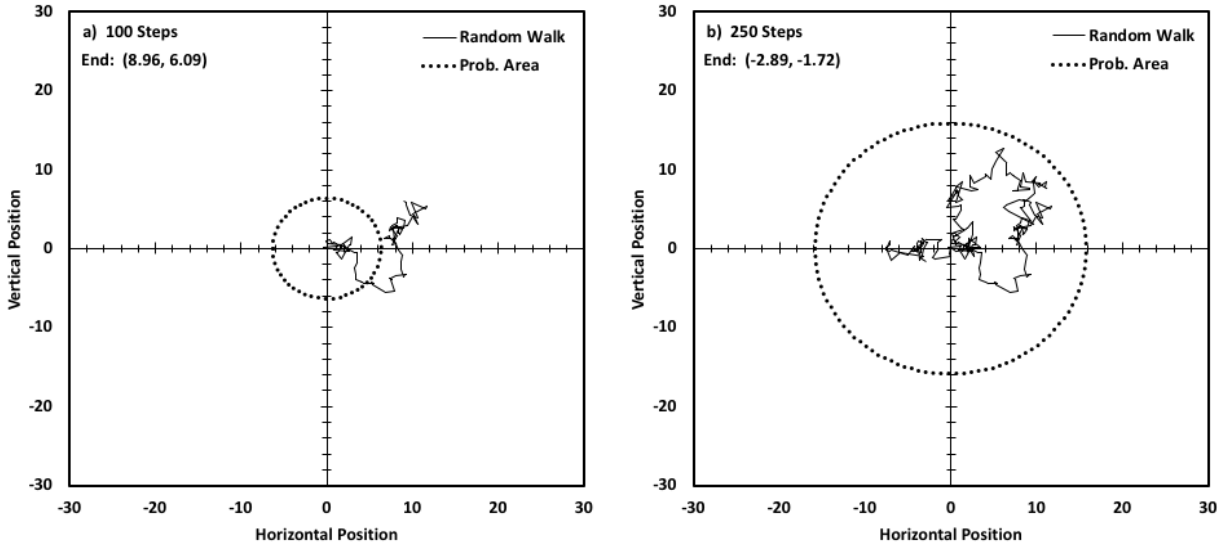

Figure C.3.1: Standard normal random walk in two-dimensions.

Figure C.3.1. illustrates the relationship between the random walk and the probabilistic area that characterizes the macroscopic process of diffusion and reflects the probabilistic area for the transition from 100 steps ($\mathcal{T} = 100$) to 250 steps ($\mathcal{T} = 250$). The radial diffusion coefficient is estimated using the standard variance of the linear displacement, $D_{r,t,n} = \frac{s_{\ell,t,n}^2}{2\pi^2 \Delta t} \implies s_{\ell,t,n}^2 = 1$. When considering a time-interval for each displacement of $\Delta t = 1$, this reflects a radial diffusion coefficient of $D_{r,t,n} = 0.0507\ \frac{Length^2}{time}$. This diffusion coefficient characterizes the time-derivative of the dispersive process for a unit source to be, $\frac{1}{\sigma_{\mathbb{A},t,n}} = \pm \frac{1}{\sqrt{\frac{\pi}{2}} D_{r,tn} \mathcal{T}} \implies \frac{1}{\sigma_{\mathbb{A},t,n}} = \pm \frac{12.37}{\mathcal{T}}$. The probabilistic areas presented on Figure C.3.1 represent the probability that the random walk will land within one standard deviation of the expected position, which is the origin because this numerical example considers a zero-centered process with no active advection.

A Random Walk with Active Advection

This section introduces advection to the stationary random process and constitutes and advective-dispersive process with respect to time. Visualization of this process may reflect diffusion of molecules within a flowing stream or even a person haphazardly attempting to walk in a straight line. The advective process becomes active within the random walk when the time-derivative of the median is assigned to be nonzero $\frac{dm_{\mathbb{A},t,n}}{dt} \neq 0$ with direction $\theta = \frac{\pi}{4}$. Changing the position of the probabilistic distribution through the median will allow the random walk to drift in a particular direction. For instance, defining the advective drift to be $\frac{dm_{\mathbb{A},t,n}}{dt} = 1$ for the same random walk presented in Figure C.3.1 produces a numerical example where the advective and dispersive processes have the same average step size. Figure C.3.2 presents the random walk from Figure C.3.1 with advective drift.

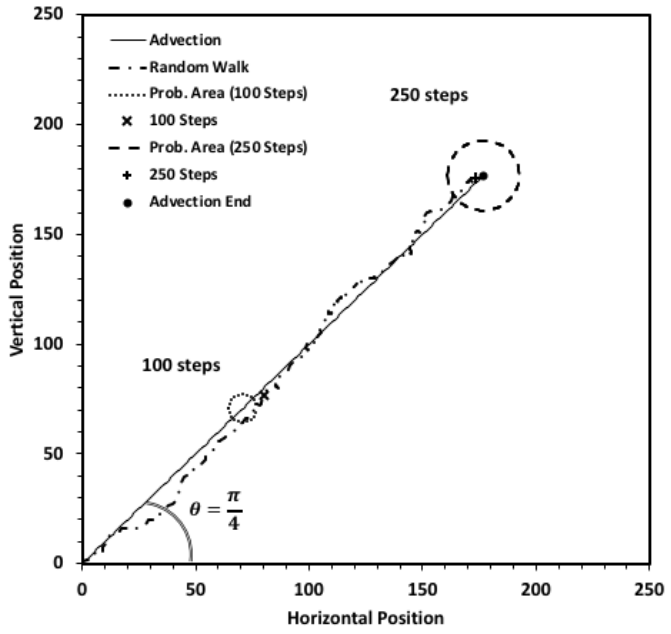

Figure C.3.2: Standard normal random walk with drift.

The advective drift term is a deterministic component of probabilistic solution that progresses the random walk in a consistent direction with a predefined step distribution. There is an associated probabilistic area for each step and the radius of this area grows with the standard deviation, $\sigma_{\mathbb{A},t,n} = \sqrt{\frac{\pi}{2}} D_{r,t,n} \mathcal{T}$. In this context, the accuracy of the advective process will be proportional to the product of the diffusion coefficient and the time-interval. For short time-intervals, the random walk is not able to stray far away from the expected position; and, a small diffusion coefficient would cause the random walk to approximate the expected position for longer time intervals. Therefore, controlling the advective drift term $\frac{dm_{\mathbb{A},t,n}}{dt}$, the diffusion coefficient $D_{r,t,n}$, and the length of time between measurements $\mathcal{T}$ could provide a predictive relationship between the random walk and the expected accuracy of the process.

Guided Random Walk

This section introduces a guided random walk, which allows the advective term to designate a direction that continually moves the distribution toward a predefined destination. Therefore, this form of random walk allows the advective term to compensate for deviation from the expected path toward the designated coordinates. Here, the progression considers a recursive

relationship between the direction of advection and some predefined destination that the advective process is attempting to achieve. The random walk can move the guided advection in a direction away from its destination, which could require the advective process to adjust its angle. This relationship can be expressed mathematically using the coordinates of the destination and the position of the random walk as:

$$\theta = ATAN\left(\frac{a_{goal} - x_t}{b_{goal} - y_t}\right)$$

Where, the direction of travel is always toward the goal position at $coor(a, b) = coor\left(a_0 + \frac{da}{dt}\Delta t, b_0 + \frac{db}{dt}\Delta t\right)$. This numerical example considers the speed of advection to be one-unit length per time interval, $\frac{dm_{\mathbb{A},t,n}}{dt} = 1$. Figure C.3.3 presents a guided random walk towards a destination of $coor(a_{goal}, b_{goal}) = coor(200, 200)$. The guided advection directs the walk to counteract the influence of the randomness moving the walk away from the destination of $coor(200, 200)$. The result is a guided random walk that always approaches the goal, while using advective drift to counteract any randomness that causes movement away from the destination coordinates.

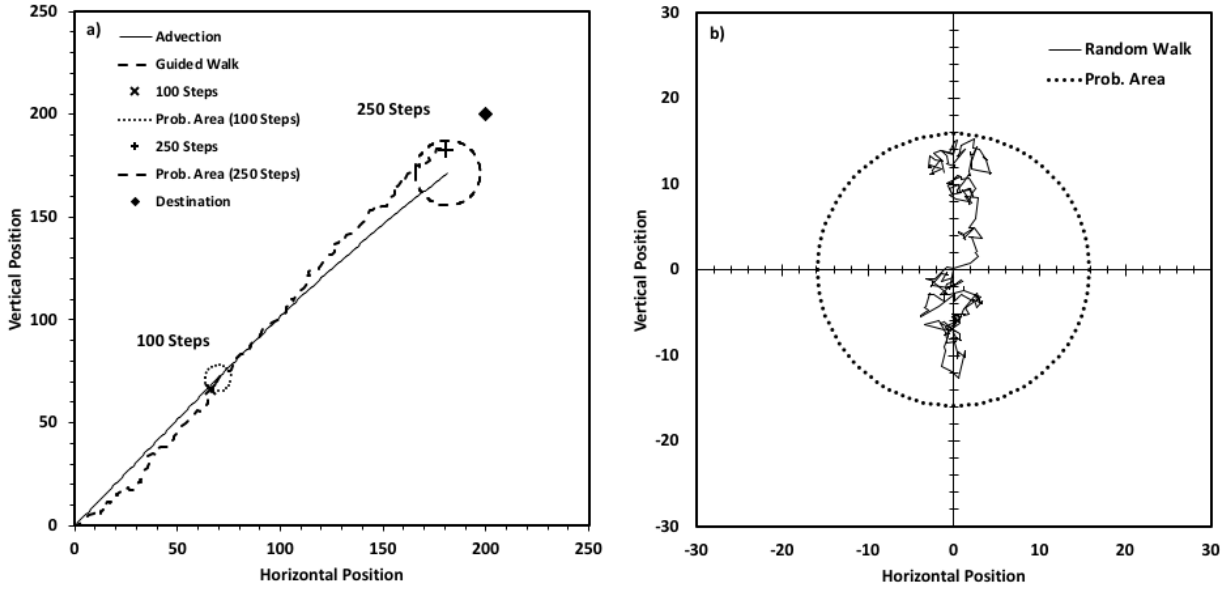

Figure C.3.3: Advection-guided, standard normal random walk in two dimensions.