12-1-2015

Compression of climate data through Artificial Neural Networks

Astha Puri
University of Nevada, Las Vegas, puri@unlv.nevada.edu

COMPRESSION OF CLIMATE DATA THROUGH ARTIFICIAL NEURAL NETWORKS

By

Astha Puri

Bachelor of Engineering
Chitkara Institute of Engineering and Technology
Punjab, India
2011

A thesis submitted in partial fulfillment
of the requirements for the

Master of Science in Engineering– Electrical Engineering

Department of Electrical and Computer Engineering
Howard R. Hughes College of Engineering
The Graduate College

University of Nevada, Las Vegas
December 2015

UNLV | GRADUATE COLLEGE

Thesis Approval

The Graduate College
The University of Nevada, Las Vegas

October 6, 2015

This thesis prepared by

Astha Puri

entitled

Compression of Climate Data through Artificial Neural Networks

is approved in partial fulfillment of the requirements for the degree of

Master of Science in Engineering – Electrical Engineering
Department of Electrical and Computer Engineering

Shahram Latifi, Ph.D.
Examination Committee Chair

Kathryn Hausbeck Korgan, Ph.D.
Graduate College Interim Dean

Henry Selvaraj, Ph.D.
Examination Committee Member

Sahjendra N. Singh, Ph.D.
Examination Committee Member

Laxmi Gewali, Ph.D.
Graduate College Faculty Representative

ABSTRACT

COMPRESSION OF CLIMATE DATA THROUGH ARTIFICIAL NEURAL NETWORKS

By

Astha Puri

Dr. Shahram Latifi, Examination Committee Chair
Professor, Electrical and Computer Engineering Department
University of Nevada, Las Vegas

Lately, there has been a tremendous increase in the number of climate monitoring stations in various parts of the country producing abundant climate data. Among climate data parameters, humidity and temperature are the two parameters influencing hydrological and agricultural processes, weather monitoring, and having critical effect on living organisms. As more data is being generated over time, there is a strong need to develop compression methods for efficient transfer and storage of this data.

The main goal of this thesis is to perform compression of humidity and temperature data via prediction. As these are critical components of climate, it is important that compression of this data is lossless.

Data for this thesis is collected from 'Nevada Climate Change Portal' (NCCP) and 'United States Geological Survey' (USGS). Humidity data comprises of 1 and 10-minute interval data for various sites in Nevada for 2013 and 2014, and temperature data comprises of hourly data from 1999 to 2012. The methodology is based on Artificial Neural Networks (ANN) to predict outputs. Feed-forward ANN model is used, in which learning is facilitated through back propagation. Modeled and observed values for humidity and temperature are compared and accuracy of the model is assessed. Differential encoding is then applied

followed by Huffman coding. This is compared with results obtained by directly applying differential encoding and Huffman coding to raw data. Performance of the method is measured by metrics like Compression Ratio (CR) and Root Mean Square Error (RMSE).

Results indicate that the predicted model gives higher compression ratio when compared to conventional method. In case of humidity, for 1-minute interval data, maximum compression ratio of 6.14 and 2.66 is achieved using proposed and conventional method respectively; and for 10-minute interval data, maximum compression ratio of 5.26 and 2.24 is achieved using proposed and conventional method respectively. In case of hourly temperature data, maximum compression ratio of 4.52 using proposed and 2.95 using conventional method is obtained. Data fluctuations being less for minute level data and values being closer to one another results in better predictions and thus higher compression ratios compared to 10-minute or hourly interval data.

ACKNOWLEDGEMENT

Although this thesis lists a single author, it is the culmination of a collaborative effort involving more people than this acknowledgement could fit. I would like to express my deepest gratitude to my advisor, Dr. Shahram Latifi, for his guidance and support for this research. His encouragement and valuable suggestions have helped me immensely in seeking the right direction for this thesis. I would also like to thank Dr. Henry Selvaraj, Dr. Sahjendra Singh and Dr. Laxmi Gewali for serving my committee and reviewing my thesis. I would also like to acknowledge NSF for providing me with the opportunity to work in the project titled "Solar Energy-Water-Environment Nexus". This work was supported by the National Science Foundation (NSF) grant #EPS-IIA-1301726.

This thesis has been a challenging experience and was accomplished with the help of many people. My deepest gratitude to my parents Shashi Kant Puri and Latika Puri for their love, care and support at every stage of my life. I would also like to thank my brother Sumit Puri and sister-in-law Surbhi Puri for their continued support and guidance in building my career. I extend my appreciation and gratitude to my fiancé Aditya Sharma for his support during my thesis. Last but not the least; I would like to thank all my friends Bharath Chandra, Sharath Shaji, Govind Pathak and Samta Garg for their support and constant encouragement throughout my Master's degree.

TABLE OF CONTENTS

ABSTRACT iii

ACKNOWLEDGEMENT v

LIST OF TABLES viii

LIST OF FIGURES x

CHAPTER 1 INTRODUCTION 1

1.1. Research Motivation 3
1.2 Research Objectives 4
1.3. Related Work 6

CHAPTER 2 BACKGROUND ON RESEARCH SITES 9

2.1 Study area 9
2.2 The Nevada Climate Change Portal (NCCP) 9
2.2.1. Research sites and equipment 11
2.2.2. Data architecture and management 17
2.2.3. Data search interface 18
2.3. The United States Geological Survey (USGS) portal 19
2.3.1. The American Drylands project 20
2.3.2. Southwest Climate Impact Meteorological Stations (CLIM-MET) 20
2.3.3. CLIMET stations 21
2.4. Summary 23

CHAPTER 3 DATA COMPRESSION 24

3.1. Introduction 24
3.2. Codes and Communication 25
3.3. Basic concepts 25
3.3.1. Compression Ratio (CR) 26
3.3.2. Types of compression 26
3.4. Lossless compression 26
3.5. Lossy compression 27
3.5.1. Scalar quantization 28
3.5.2. Vector quantization 29
3.6. Differential encoding 29
3.7. Huffman coding 30
3.7.1. Compression 31
3.7.2. Decompression 31
3.8. Summary 32

CHAPTER 4 ARTIFICIAL NEURAL NETWORKS 33

4.1. Introduction 33
4.2. Model of a neuron 34
4.3. Characteristics of neural networks 35
4.4. Learning process 36
4.4.1. Supervised learning 36
4.4.2. Unsupervised learning 37
4.4.3. Reinforced learning 38
4.5. Feed forward network 38
4.6. Multi layer perceptron (MLP) 39
4.7. Back propagation algorithm 40
4.7.1. Phase 1: Propagation 41
4.7.2. Phase 2: Weight update 41
4.8. Benefits of neural networks 42
4.9. Summary 43
CHAPTER 5 DATA COMPRESSION USING ARTIFICIAL NEURAL NETWORKS 44
5.1. Overview 44
5.2. Parameters studied 44
5.2.1. Humidity dataset 44
5.2.2. Temperature dataset 45
5.3. Sites considered 45
5.3.1. Snake Range West Subalpine 45
5.3.2. Sheep Range Mojave Desert Shrub 46
5.3.3. North Soda Lake 46
5.4. Artificial Neural Network (ANN) Model 46
5.4.1. Humidity model 47
5.4.2. Temperature model 49
5.5. Proposed method 51
5.5.1. Method 1 51
5.5.2. Method 2 54
5.6. Results 54
5.6.1. Results of humidity data 55
5.6.2. Results of temperature data 62
CHAPTER 6 CONCLUSION AND RECOMMENDATIONS 66
6.1. Conclusions 66
6.2. Recommendations 67
6.3. Limitations 68
REFERENCES 69
CURRICULUM VITAE 74

LIST OF TABLES

Table 1. Site descriptions of locations along the two transects comprising the NevCAN.....13

Table 2. List of Sensors.....17

Table 3. List of equipment employed at the CLIMET sites.....23

Table 4. Observed parameters and units of measurement for humidity model. 45

Table 5. Observed parameters and their units of measurement for temperature model. 45

Table 6. Details of Snake Range West Subalpine. 46

Table 7. Details of Sheep Range Mojave Desert Shrub.....46

Table 8. Details of North Soda Lake.....46

Table 9. Input and output parameters for humidity model.....48

Table 10. Input and output parameters for temperature model.....50

Table 11. MSE and RMSE vales for Sheep Range Mojave Desert Shrub.....55

Table 12. Monthly compression ratios for the year 2014 for minute interval humidity dataset for Sheep Range Mojave Desert Shrub.....56

Table 13. Monthly compression ratios for the year 2014 for 10-minute humidity for Sheep Range Mojave Desert Shrub.....57

Table 14. MSE and RMSE values for Snake Range West Subalpine. 58

Table 15. Monthly compression ratios for the year 2014 for 1-minute humidity data for Snake Range West Subalpine.....60

Table 16. Monthly compression ratios for the year 2014 for 10-minute humidity for Snake Range West Subalpine.....61

Table 17. Monthly compression ratios for the year 2012 for temperature data for North Soda

Lake.....65

LIST OF FIGURES

Figure 1. NCCP's landing page provides quick access to all its offerings. 10

Figure 2. Location of the EPSCoR – NevCAN stations, one transect of 8 stations traverses the Snake mountain range and another transect of 5 stations is situated on the Sheep Range 12

Figure 3. The "Sensors and Equipment" page shows users the hardware deployed on the field. 14

Figure 4. List of Web Cameras at each location of NevCAN stations. 15

Figure 5. Cross sectional schematic of standard instrumentation installed at each transect site. 16

Figure 6. The Data Search Interface allows the user to query and download available climate data 18

Figure 7. The location of CLIM-MET stations on a US map. 21

Figure 8. Equipment placed at the CLIMET stations 22

Figure 9. Lossless compression. 27

Figure 10. Sinusoid and sample to sample difference. 30

Figure 11. An artificial neuron. 34

Figure 12. Non-linear model of a neuron 34

Figure 13. Supervised learning. 37

Figure 14. Unsupervised learning 38

Figure 15. Feed forward neural network design. 39

Figure 16. A fully connected multilayer feedforward network 40

Figure 17. ANN Network model for 1-minute data for Snake Range West Subalpine. 48

Figure 18. ANN Network model for 10-minute data for Snake Range West Subalpine. 48

Figure 19. ANN Network model for 1-minute data for Sheep Range Mojave Desert Shrub. .49

Figure 20. ANN Network model for 10-minute data for Sheep Range Mojave Desert Shrub 49

Figure 21. ANN model for temperature data. 50

Figure 22. Flowchart for compression algorithm. 51

Figure 23. Flowchart for decompression algorithm.....52

Figure 24. Actual and predicted humidity values for Sheep Range Mojave Desert Shrub. 55

Figure 25. Comparison of compression ratio obtained with two methods for minute interval dataset of Sheep Range Mojave Desert Shrub.....57

Figure 26. Comparison of compression ratio obtained with two methods for 10-minute interval dataset of Sheep Range Mojave Desert Shrub.....58

Figure 27. Actual and predicted humidity values for minute interval dataset for Snake Range West Subalpine.....59

Figure 28. Comparison of compression ratio obtained with two methods for minute interval dataset of Snake Range West Subalpine. 61

Figure 29. Comparison of compression ratio obtained with two methods for 10-minute interval dataset of Snake Range West Subalpine. 62

Figure 30. Actual and predicted temperature data for North Soda Lake, California. 63

Figure 31. Comparison of compression ratio obtained with two methods for temperature data. 64

CHAPTER 1

INTRODUCTION

In today's world, data compression is ubiquitous. It is now a vital part of everyday life, be it for text, images, audio or video. Images on the web are all compressed, typically in JPEG formats; Modems use compression; Several systems compress files when storing them [1]. An aspect of data compression that should be examined is the enormous amount of climate data which requires effective compression techniques. This thesis will address the above issue using Artificial Neural Networks along with compression algorithms.

The need for development of more efficient ways of representing information increases with the increase in the amount of information that is needed, desired, and available. The goal of data compression is to provide most efficient ways of representing information. To accomplish this goal, different techniques have been developed to exploit various kinds of structures that may be present in the data. Information can be in various forms, such as speech, images, text, video and so on. Each form of information has its own specific types of structures. These forms also share characteristics that can often be exploited to develop techniques that have relevance to all kinds of information [2].

Data compression is the art and science of representing information in a compact form, i.e., reducing the number of bits needed to store or transmit data. It is sometimes referred to as source coding or bit-rate reduction. Data can be characters in a text file, decimal data, numbers that are samples of speech or image waveforms, or a sequence of numbers that are generated by different processes. Compact representation of this data can be created by identifying and exploiting structures that exist in this data [2]. The task of compression requires an encoding algorithm that takes a message and generates a "compressed" representation (with fewer bits), and the decompression task requires a decoding algorithm

that reconstructs the original message or some approximation of it from the compressed representation [1].

An early example of data compression is Morse code, invented in 1838 for use in telegraphy. It is based on using shorter code words for letters that are common in the English language. Morse noticed that frequency of occurrence of some letters was more than other letters. He assigned shorter sequences to letters that occurred more frequently, such as "a" and "e," which reduced the average time required to send a message. Another example is Braille coding, in which text is represented using 2 x 3 arrays of dots. Each letter is represented by different combination of raised and flat dots [1].

Modern work on data compression began in the late 1940s with the development of information theory. In 1949, a systematic way to assign code words based on probabilities of blocks was devised by Claude Shannon and Robert Fano. This was further optimized by David Huffman in 1951. For Huffman coding, in the mid 1970s, an idea emerged for dynamically updating code words based on the actual data encountered. Late 1970s saw the online storage of text files becoming common, which led to the development of software compression programs, almost all based on adaptive Huffman coding. Abraham Lempel and Jacob Ziv suggested the basic idea of pointer-based encoding in 1977. In the mid 1980s, Terry Welch followed this work and the LZW algorithm became the choice for most compression systems. It was used in hardware devices, such as modems, as well as in programs such as PKZIP [3].

With the explosive development of internet, mobile and video communications, compression has become necessary. Compression plays an important role in the ability to transmit digital television signals. It if were required to transmit HDTV signal without compression, 884 Mbits would need to be transmitted per second over a 220 MHz channel. However, with data compression, less than 20 Mbits per second would be needed for

transmission, for which only a 6 MHz bandwidth channel is needed. Fax machines also use data compression; without compression, it would take 24 hours to send a 180-page document. As we generate and use more information in digital form, which consists of numbers represented by bytes of data, bytes required to represent multimedia data can be huge. For example, in order to digitally represent one second of video without compression would require more than 20 megabytes of data [2].

Various space agencies from around the world, such as, the National Aeronautics and Space Administration (NASA), are collaborating on a program that will generate half a terabyte of data per day when it is operational. Therefore, it is imperative to develop compression techniques that reduce the amount of storage space required.

Data compression reduces data, so it requires less disk space for storage and less bandwidth on a data transmission channel. Communication equipment like modems, routers etc; use compression schemes to improve performance and to provide communication with increased clarity. Other applications where compression is used are voice telephone calls, videoconferencing, and audio recordings, among others. Compression helps in faster reading and writing, easier and faster file transfer, and saving disk space.

Thus, data compression, being an important, necessary and critical field of study, requires efficient ways to compress data and will be studied and developed in this thesis.

1.1. Research Motivation

Weather data comes from diverse sources. It can be obtained from human reports, in situ instruments, or remote sensors. Data produced from meteorological and climatological networks and various research projects represents a valuable and unique resource, costing a substantial amount of time, money and effort. The initial use of meteorological and related data for weather forecasting is often only the first of many applications. Subsequent analysis of data for diverse purposes leads to a significant and ongoing enhancement in data

management. The global climate change issue, for example, is stretching the need for climate data and data management systems.

The main interest in the use of observed climatological data is not to simply describe the data, but to make inferences that are helpful to users of climatological information. Statistics from the observed data is used to make such inferences. A tool called "Statistics" which is used to bridge the gap between raw and useful information can be employed for analyzing data and climate prediction models. For example, statistics are used to identify trends in climate, such as precipitation days [4].

The importance of climate data cannot be ignored. Forecasters use meteorological data to support a number of programs including public, aviation, fire, and marine. Forecasters preparing public products routinely monitor temperature data to produce 1 to 7 day forecasts of temperature. Aviation forecasters keep an eye on surface observations for wind sheer, weather, or restrictions to visibility that could adversely affect take-offs and landings. Data associated with events that can result in loss of life and destruction of property, such as severe storms, hurricanes, and extreme winter storms, must be carefully monitored in order to issue timely and accurate warnings [3].

Climate databases are enormous in size and extremely important for research studies. Since climate data is used in almost every field, it is important that data be accurate and usable. Recently, there have been issues of space for data storage since, on an average, gigabytes of climate data is being generated and stored in climate databases throughout the world. To manage this voluminous data, efficient methods for storage and transmission should be considered. In this study, compression of humidity and temperature data is addressed.

1.2 Research Objectives

This research aims to develop a model that can calculate the compression ratio of

humidity and temperature data. Humidity data used is from Nevada Climate Data Portal (NCCP) and temperature data used is from United States Geological Survey (USGS). Humidity, which is a measure of amount of moisture in the air, has a large impact on human and animal health as well as the health of crops. For this study, humidity data is acquired from sensors placed in the field. CSI HMP50 sensor is used to collect the data. Air temperature, which is a measure of thermal or internal energy of molecules within an object or gas, affects oceans, weather patterns, snow and ice, as well as plants and animals. Higher the air temperature, more severe the impact on the people and environment. Temperature can be measured by direct contact or remote sensing. For this study, temperature data is observed from sensors placed at specific locations. As both these parameters, i.e. humidity and temperature, are critical components of climate, it is important that compression of this data be done accurately.

The objective of this thesis is to use artificial neural networks for predictive analysis, i.e., to predict climate data, and then combine predicted outputs with lossless compression algorithm, i.e., differential encoding followed by Huffman coding, to calculate improved results for compression ratios.

Key research questions addressed in this study are as follows:

1. How compression ratio varies with and without predictive analysis method?
2. How results vary with the location of site?
3. What is the maximum compression ratio observed for each parameter?

To address these questions, following tasks were undertaken:

1. Humidity data from Nevada Climate Change Portal and temperature data from the United States Geological Survey are obtained.
2. Using artificial neural networks, a model is developed for each parameter.
3. To train the neural network, humidity data for the year 2013 is used and temperature data

for the years 1999 to 2011 is used to obtain the modeled values.

4. The modeled values of humidity and temperature, and their observed values, are compared and the accuracy of the model is assessed.

5. The modeled, i.e., the predicted values, are then used to perform differential encoding, which is followed by Huffman coding.

6. The performance of prediction is calculated in terms of root mean square error and finally, the compression ratio is calculated.

7. The process is repeated for different sites under study.

This research is organized as follows: Chapter 2 presents background information on the area of study; Chapter 3 gives an introduction to data compression; Chapter 4 describes Artificial Neural Networks; Chapter 5 gives the results, and Chapter 6 concludes this study.

1.3. Related Work

This study focuses on lossless compression of climate data. A technique is presented in [5] to compress pressure, wind and precipitation data, where uniform quantization is applied to pressure data as part of preprocessing. Optimal prediction techniques are used to predict the values based on the surrounding values and the differences are encoded. Wind data, which includes wind velocity and wind direction, are transformed into polar co-ordinate form. Entropy coding is performed after the preprocessing stage. Also, significant improvements in bandwidth can be realized through the use of common compression techniques, such as wavelet/error grid or round/difference/BZIP2 compression [6]. The compression of temperature data in Kuala Lampur from January 1948 until July 2010 by using Debauchies wavelet (D4) as the basis function is performed in [7]. Various approaches for scientific data compression have focused primarily on combining compression with data synthesis in order to increase throughput and conserve storage. In [8] Engelson compressed sequences of double precision floating point values resulting from simulations based on

ordinary differential equations. In a theoretical approach, the numbers are treated as integers and then compressed using predictive coding, with residuals being explicitly stored in case of lossless coding or truncated for lossy coding. In [9], a lossless prediction based compression method for double-precision floating point scientific data is proposed using a DFCM (Differential Finite Context Method) value method, which is based on pattern matching using a hash table holding recent encoding context. The bitwise residual was then computed using XOR operator, with compressed representation consisting of a finite number of leading zeros and remaining residual bits.

The Artificial Neural Network models use different geographical parameters of a location as inputs for the prediction of solar radiation as discussed in [10]. In [11], a multi-layer feed forward network is discussed, and a back propagation (BP) training algorithm for global radiation prediction is used. The area of study is Seeb, Oman. The inputs used in the network were location, month, mean pressure, mean temperature, mean vapor pressure, mean relative humidity, mean wind speed and mean sunshine hours. In the study by Sözen et al. [12] [13], meteorological and geographical data was used as input variables for the ANN model for solar radiation estimation in Turkey. The transfer function for this model is logistic sigmoid and the learning algorithm is Scaled conjugate gradient, Pola-Ribiere conjugate gradient Levenberg-Marquardt. In the study undertaken in [14], back propagation neural network (BPNN) models were developed to predict the day soil temperature for the present day by using various previous-day meteorological variables in Nineveh-Iraq. Three models were developed, the first using back propagation, the second using Cascade back propagation, and the third using the NARX, which consisted of the combination of the input variables. The models were constructed to obtain the best fit input structure. After ANN training; 75%, 80% and 95% of the data points were the same as actual data in the first, second and third model respectively. In [15], the problem of lossless, offline compression of

climate data was addressed. A method for the compression of solar radiation, photo synthetically active radiation, and data logger power system voltage data using a combination of differential encoding and Huffman coding was developed.

CHAPTER 2

BACKGROUND ON RESEARCH SITES

This chapter is organized as follows: Section 2.1 discusses the study area and its details; Section 2.2 describes the 'Nevada Climate Change Portal' and its methods used for collecting data; Section 2.3 discusses the 'United States Geological Survey' portal, and Section 2.4 gives the summary.

2.1 Study area

There are two areas of study in this research. The humidity parameter is studied from the 'Nevada Climate Change Portal' over two sites in the state of Nevada, and the temperature parameter is studied from the 'United States Geological Survey' portal for the sites in California. The characteristics of each study area are described in the sections below.

2.2 The Nevada Climate Change Portal (NCCP)

Under the NSF EPSCoR-funded RII (Track I) project, the Spatial Engine for Nevada Scientific Observational Results (SENSOR) system was created to provide an automated mechanism for long-term storage, acquisition, curation and management of raw observational research data. Software systems, modern hardware and an advanced geospatial database system are combined to provide high-quality, lossless data streams and auto-generated metadata for researchers, students, and public [16].

One of the key products that came out of an NSF EPSCoR RII Track 1 project entitled “Nevada Infrastructure for Climate Change Science, Education, and Outreach” was the Nevada Climate Change Portal (NCCP). It is designed for data acquisition, storage, access, and processing in support of the long-term assessment of climate variability in Nevada and its impact on the state’s ecological and hydrological systems. It was the result of a collaborative effort amongst researchers, disciplines, and institutions. It serves as a cyber-infrastructure hub

and a central repository of climate-related information for other stakeholders, including educators, students, and public. It provides data and computing resources for scientists studying the effects of climate change in Nevada. NCCP has various components including data resources, retrieval and processing tools, software solutions for facilitating scientific research, research sites, people involved, tutorials and publication/photo/video libraries [17].

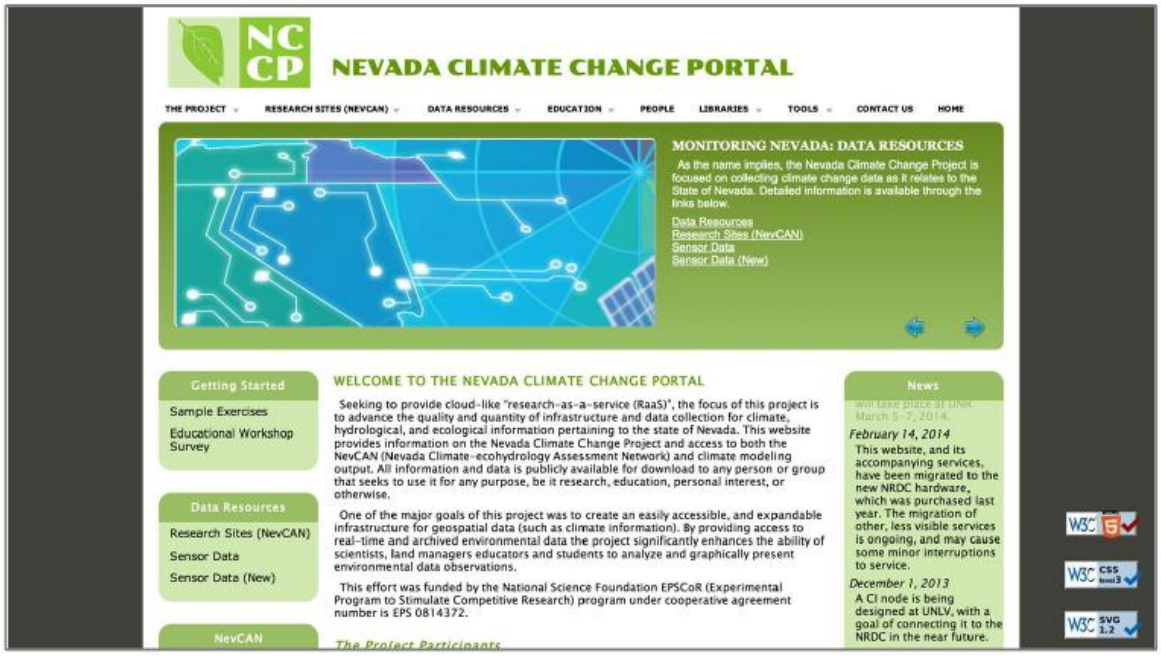

Figure 1. NCCP's landing page provides quick access to all its offerings [18].

As per the April 2014 update, the NCCP has collected over 1.2 billion high-quality environmental measurements and this number is increasing each day. There are thirteen Nevada Climate-ecohydrological Assessment Network (NevCAN) research sites spread across the state; these provide scientists and other researchers with measurements of air, water, plant and soil readings, for numerous long and short-term analyses.

Data collected by NCCP is freely available at http://sensor.nevada.edu. Data is comprehensible, well managed and readily available to all interested researchers, organizations and members of the public [17].

2.2.1. Research sites and equipment

The National Science Foundation EPSCoR (Experimental Program to Stimulate Competitive Research) program in Nevada funded the establishment and construction of two elevational transects of monitoring stations named the Nevada Climate-ecohydrology Assessment Network (NevCAN). The main purpose of NevCAN is to collect data for long-term assessment of climate variability and change in Nevada, and to monitor the changes in ecological and hydrological processes. These stations were built across the elevational gradient in the Snake and Sheep mountain ranges, and all contain similar instrument packages for purposes of elevational and latitudinal comparisons of climatic, ecologic, and hydrologic variables. The Snake Range transect, located approximately 335 km NNE of Las Vegas, has eight monitoring stations beginning at 1790 m on the west side of the range, 3355 m at the western subalpine site and ending at 1560 m on the eastern side of the range. The Sheep Range transect, located approximately 35 km NNW of Las Vegas, has five monitoring stations beginning at 900 m and ending at 3015 m [19].

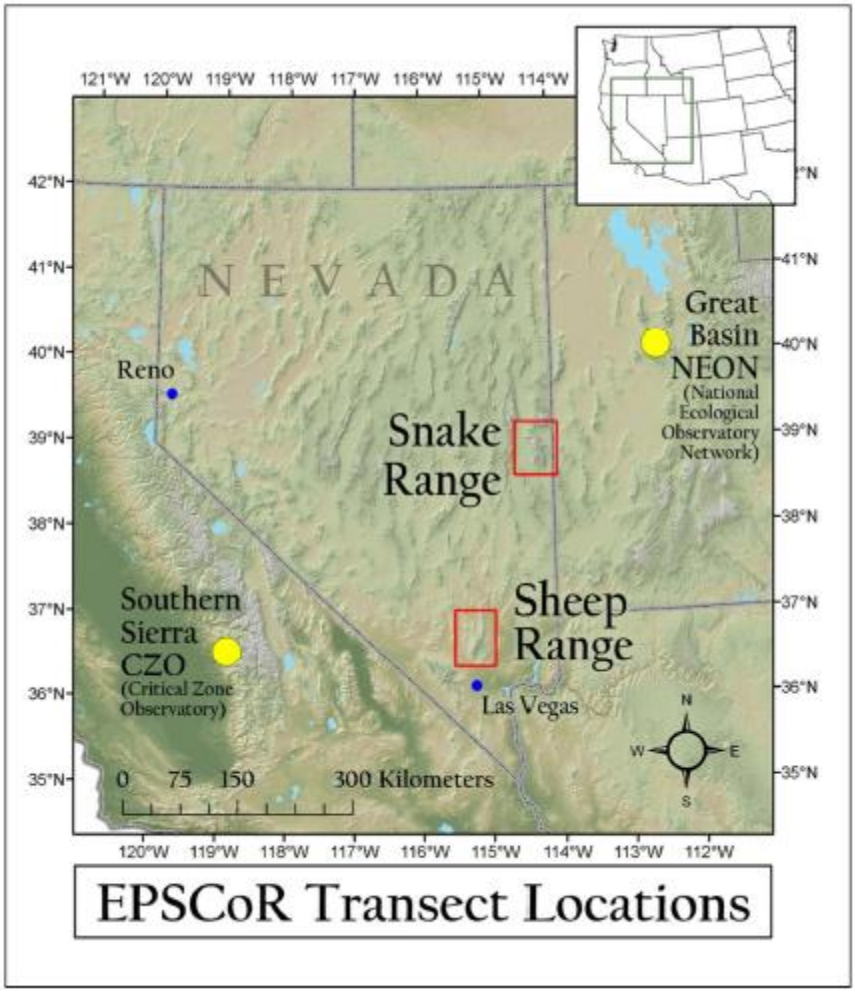

Figure 2. Location of the EPSCoR – NevCAN stations, one transect of 8 stations traverses the Snake mountain range and another transect of 5 stations is situated on the Sheep Range [17].

The area for each monitoring station is an approximate 100 x 100 m (one hectare) homogeneous area. These stations are capable of capturing variability in the winter dominated climate of the Snake Range as well as the bi-modal winter and summer monsoon-driven regime in the Sheep Range [20].

Table 1. Site descriptions of locations along the two transects comprising the NevCAN [19].

Transect	Zone	Dominant Plant Species	Altitude (m)
Snake	Salt Desert (west)	*Sarcobatus vermiculatus*	1755
Snake	Sagebrush (west)	*Artemisia tridentata tridentata*	1790
Snake	Pinyon-Juniper (west)	*Pinus monophylla*	2200
Snake	Montane (west)	*Abies concolor*	2810
Snake	Subalpine (west)	*Pinus longaeva*	3355
Snake	Subalpine (east)	*Picea engelmannii*	3070
Snake	Sagebrush (east)	*Artemisia tridentata tridentata*	1835
Snake	Salt Desert (east)	*Sarcobatus vermiculatus*	1560
Sheep	Desert Shrub	*Larrea tridentata*	900
Sheep	Blackbrush	*Yucca brevifolia*	1670
Sheep	Pinyon-Juniper	*Pinus monophylla*	2065
Sheep	Montane	*Pinus ponderosa*	2320
Sheep	Subalpine	*Pinus longaeva*	3015

Included on the NevCAN stations in the Snake Range are atmospheric/meteorological sensing systems comprised of nine different physical sensors, including free air temperature at 10m and 2m heights, relative humidity, air pressure, incoming and outgoing long wave and shortwave solar radiation, wind speed/direction, and snow depth. There are two different sensors to measure precipitation at the higher elevations, which can differentiate rain vs. snow. Water content and soil conditions are also monitored at all stations using nine sensors, which include temperature and water content in a vertical array to a depth of 50 cm. The data logging of these variables is available in the frequency of every one minute, which allows examination of the processes across landscapes on short

timescales. Additionally, there are some sites which are used for experimental purposes and deploy sensors ranging between 2-48 in size. They also measure variables like tree sap flow, snow water equivalent, distributed soil moisture and temperature, incremental tree growth, and Normalized Differential Vegetation Index (NDVI). Each year the number of sensors being added is increasing. The NevCAN sites supports 240 "standard" sensors and 150 "experimental" sensors [17].

Figure 3. The "Sensors and Equipment" page shows users the hardware deployed on the field [17].

For visualization, a controllable Point-Tilt-Zoom (PTZ) camera is installed in every station. It can capture up to twenty images per hour automatically. Each camera on every location captures between 100 and 160 images a day, adding to approximately 1300 images per day. Currently, the image collection of the NCCP contains over 1.57 million individual photographs [17].

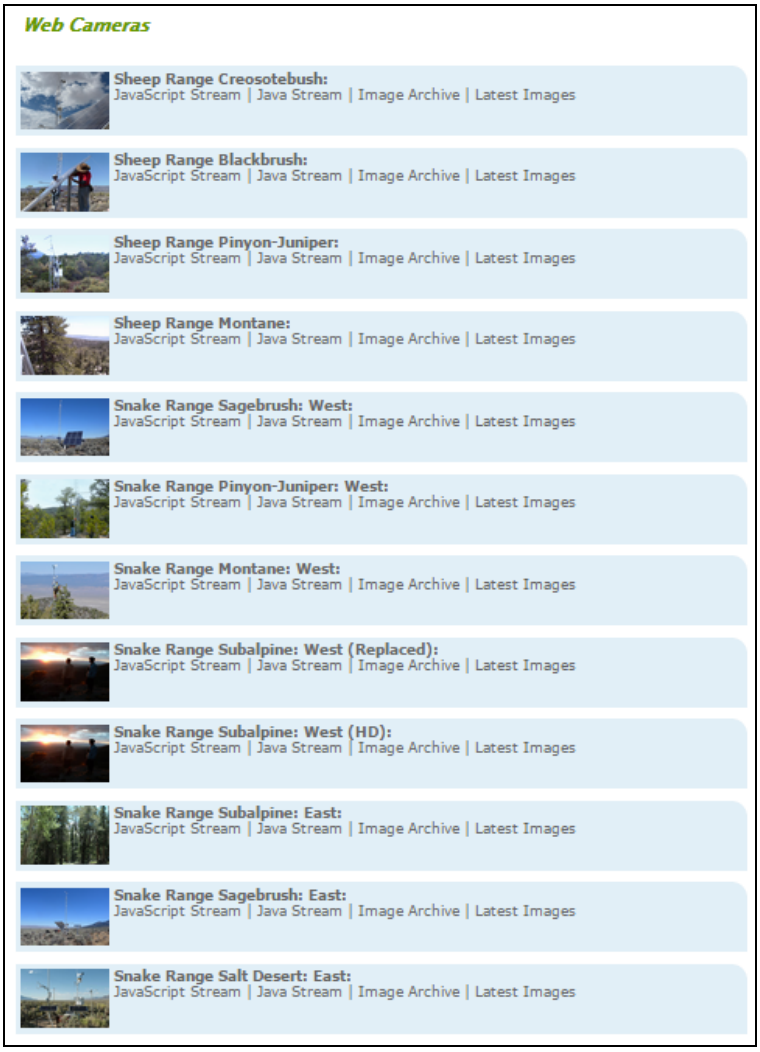

Figure 4. List of Web Cameras at each location of NevCAN stations [17].

Long-distance terrestrial wireless networking is employed for the connection between NevCAN stations and real-time control of the field devices. The servers for the connection of data loggers and cameras are located at the University of Nevada, Reno. Hundreds of kilometers of data radio links are used. Effective measures, such as parallel links, are provided to ensure connectivity in the event of interference or failure. To manage the data efficiently, regular database backups are taken [17].

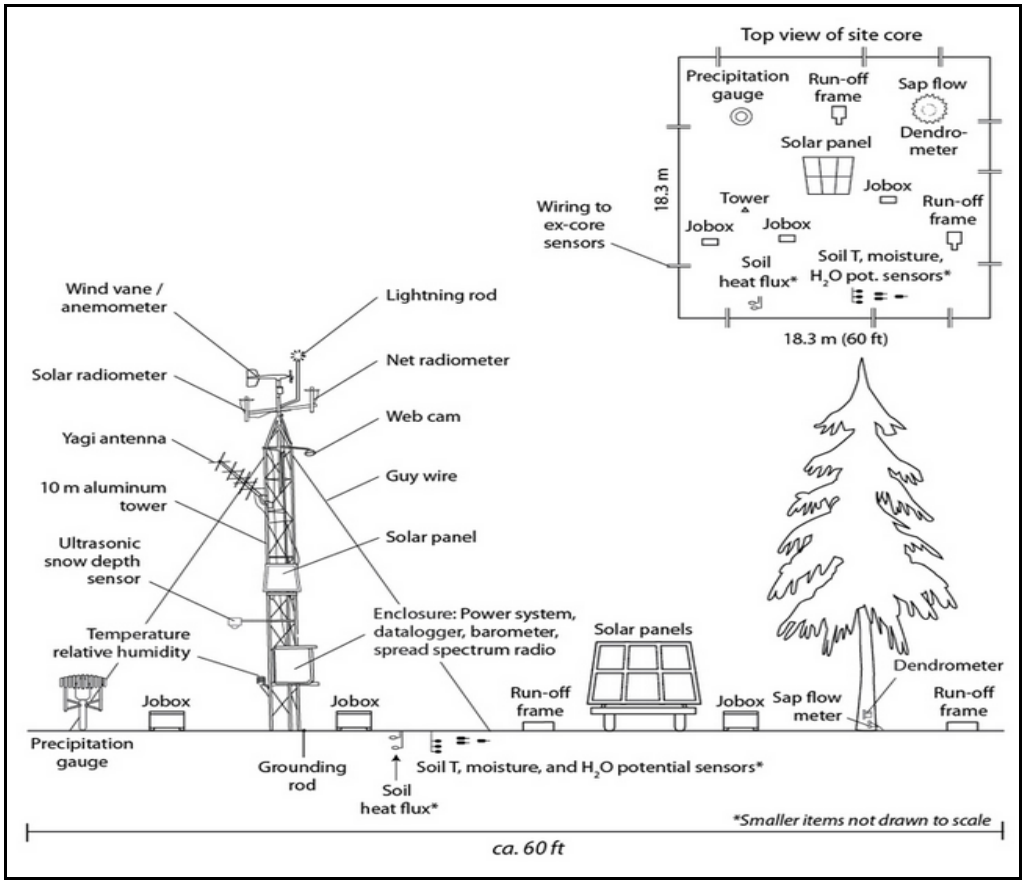

Figure 5. Cross sectional schematic of standard instrumentation installed at each transect site [17].

Climate data collected by means of these NevCAN sites across Nevada for the past four years can be accessed through NCCP. Data is transferred via radio link to a base station at Great Basin College or to base stations close to the Sheep Range sites over a frequency of 900MHz. The data portal constructed by EPSCoR's cyber infrastructure group receives all data. General public can access the real time and archived data through this data portal [21].

Table 2. List of Sensors [22].

Sensor name	Parameter measured
Ambient air temperature thermocouple (OMEGA Copper Const.)	Sensor measures air temperature.
Quantum sensor (LiCor 190SA)	Sensor measures photosynthetically active radiation.
Capacitative RH (CSI HMP50)	Sensor measures air temperature and relative humidity.
Propeller anemometer (Wind vane, RM Young 05103)	Sensor measures wind speed and direction.
Constantan thermocouple	Sensor measures soil temperature (installed at 3 depths).
Dual probe heat pulse (DPHP) sensor	Sensor measures soil thermal conductivity, diffusivity, and specific heat.
Acclima TDT monitoring system	Sensors measure soil moisture, salinity and temperature.
Pyranometer (LiCor 200SZ)	Sensor measures solar radiation.
Snow stake	Sensor measures snow depth using the webcam.

2.2.2. Data architecture and management

There are various components essential to NCCP's infrastructure. Each component is responsible for fulfilling a specific functionality and is required to interact with related components via clearly defined mechanisms. All components are based on the basic principle of preservation of high-quality measurements and associated metadata. Measurements are obtained from the remote sensing equipment through a data collection component and are then passed to the data storage component, which is a fault-tolerant redundant storage system.

The data import layer consists of software services that examine newly obtained measurement files exposed by the data storage component. The services verify the integrity and format of files as and when they arrive, and decompose them for submission to the database component. Additionally, services handle rejected submissions, manage rejected and submitted files and ensure the import process runs efficiently.

The database component is comprised of a database which forms the heart of the

system. It accepts measurements from the data import layer and integrates new files into its collection and performs any conversions if needed. In case submissions cannot be combined, a failure notification is issued to the data import layer. The data access component serves as a security layer and provides structured access to the database component ensuring data is never modified by external entities. The last component is the data curation component which operates parallel to data storage and data import components and its main purpose is to ensure the viability and quality of collected files. It also archives the raw imported files which are of importance to researchers [17].

2.2.3. Data search interface

One of the major features offered by NCCP is the Data Search Interface, through which users can search and visualize the climate data. Figure 6 shows the data search interface.

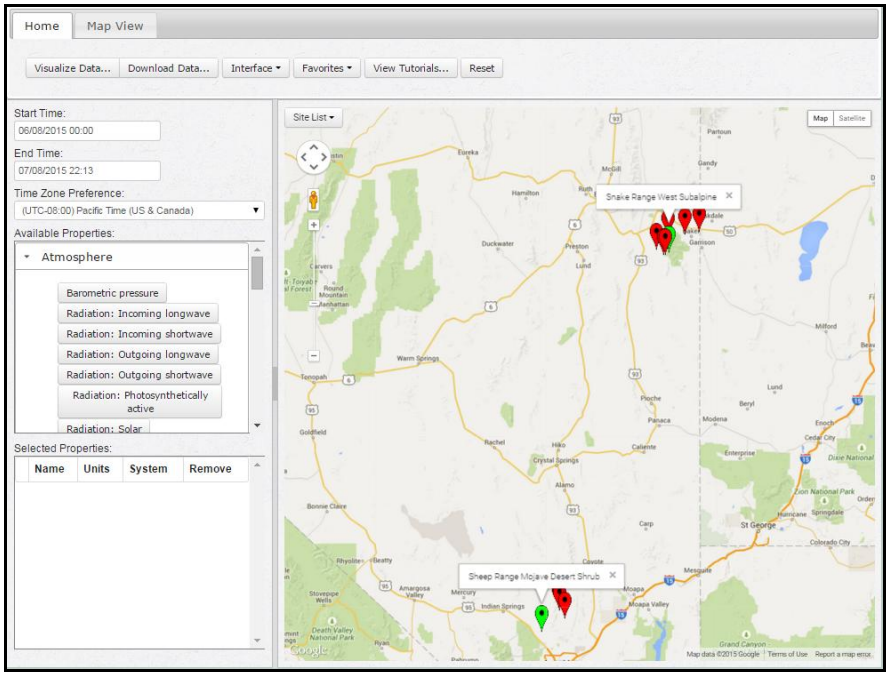

Figure 6. The Data Search Interface allows the user to query and download available climate data [23].

This interface provides quick and easy access to climate data by allowing users to query the required datasets. It has a map based site selection feature. Each site has numerous properties and sensors which provide data at a variety of intervals. This enables users to have full control over their data set; for example, specifying units for their measurement, create and run queries against the data set and visualize results [17].

2.3. The United States Geological Survey (USGS) portal

The USGS is one of the organizations providing fair information about ecosystems and environment. It provides awareness about the threats from natural hazards, impact of land-use and climate change, and science systems for providing relevant, timely and useful information.

The USGS has the following functions:

- Describe and understand the Earth.
- Minimize loss of life and property from natural disasters.
- Manage water, biological, energy and mineral resources.
- Enhance and protect quality of life.

It is responsible for collecting, monitoring, analyzing, and providing scientific understanding of natural resource conditions, problems, and issues.

One of the key goals of the USGS is to "understand past, present, and future environmental consequences of land change to support better management of their effect on people, environment, economy, and resources." To support this, scientists constantly conduct research and develop capabilities to assess climate, landscape and environmental changes. They also work to investigate ecological disturbance patterns which result from natural changes [24].

2.3.1. The American Drylands project

The American Drylands project addresses the urgent need to understand and measure physical landscape change which influences ecosystems and human. Ecological services such as water, productivity, and landscape stability are a few basic human requirements. This project aims to develop a new understanding of interactions among physical and human systems and their responses to climate change. For this purpose, changes in physical and ecological landscapes are noted.

The objectives of this project are to:

- Examine interactions among geologic, hydrologic, biologic, and atmospheric processes in drylands.
- Measure land surface and ecological changes in response to climate and human activity in several ecologically sensitive drylands.
- Forecast responses of dry landscapes to climatic variability using newly developed and established geologic, hydrologic, and biogeochemical models of landscape processes [25].

As a part of this project, CLIM-MET meteorological stations are implemented, which provide understanding on how climate, such as temperature, precipitation, wind direction and speed, and human activities affect geologic processes [26].

2.3.2. Southwest Climate Impact Meteorological Stations (CLIM-MET)

The CLIM-MET stations, which are meteorological/geological stations, are primarily designed to function in remote areas. They are created to operate for long periods of time without human involvement. The meteorological parameters are measured by these stations and data is automatically recorded at regular intervals. A system of radio-telemetry and satellite internet is used to retrieve the data in real time for the Mojave stations, whereas, for Canyonland stations, data is stored on site and retrieved four times a year. The data is

checked for validity and post calibrations and then is available on their website, which allows access to data for scientists and researchers. Inputs to regional climatic models are provided by CLIM-MET data and can also respond to the future climatic conditions. Remote sensing technique is used to capture the data.

Each station has data sets of meteorological observations, landscape responses, and hydro climatology and photographs of landscape change. The Mojave National Preserve, California location consists of three stations; namely, North Soda Lake, Balch, and Crucero. Canyonlands National Park, located in Utah contains four stations; namely, Virginia Park, Needles, Dugout Ranch, and Corral Pocket [26]. Figure 7 shows the location of these stations on the map.

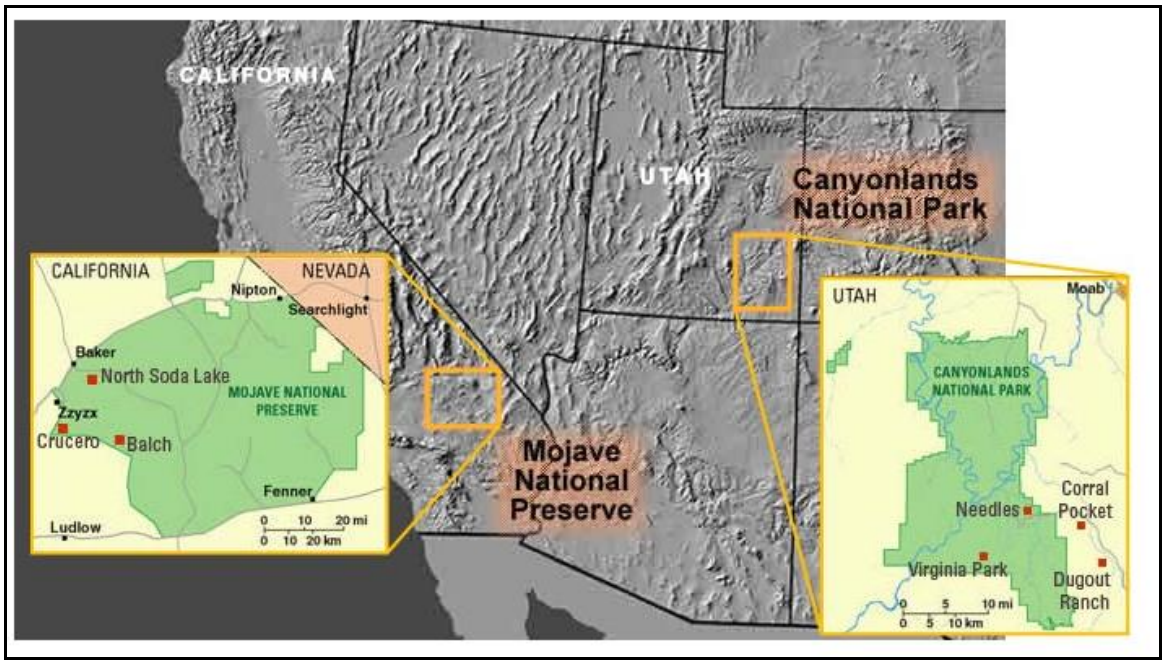

Figure 7. The location of CLIM-MET stations on a US map [26].

2.3.3. CLIMET stations

Each CLIMET stations is equipped to record data points. Table 3 gives a list of

equipment employed at these locations to collect data from sensors, and Figure 8 gives the location of this equipment on the sites graphically.

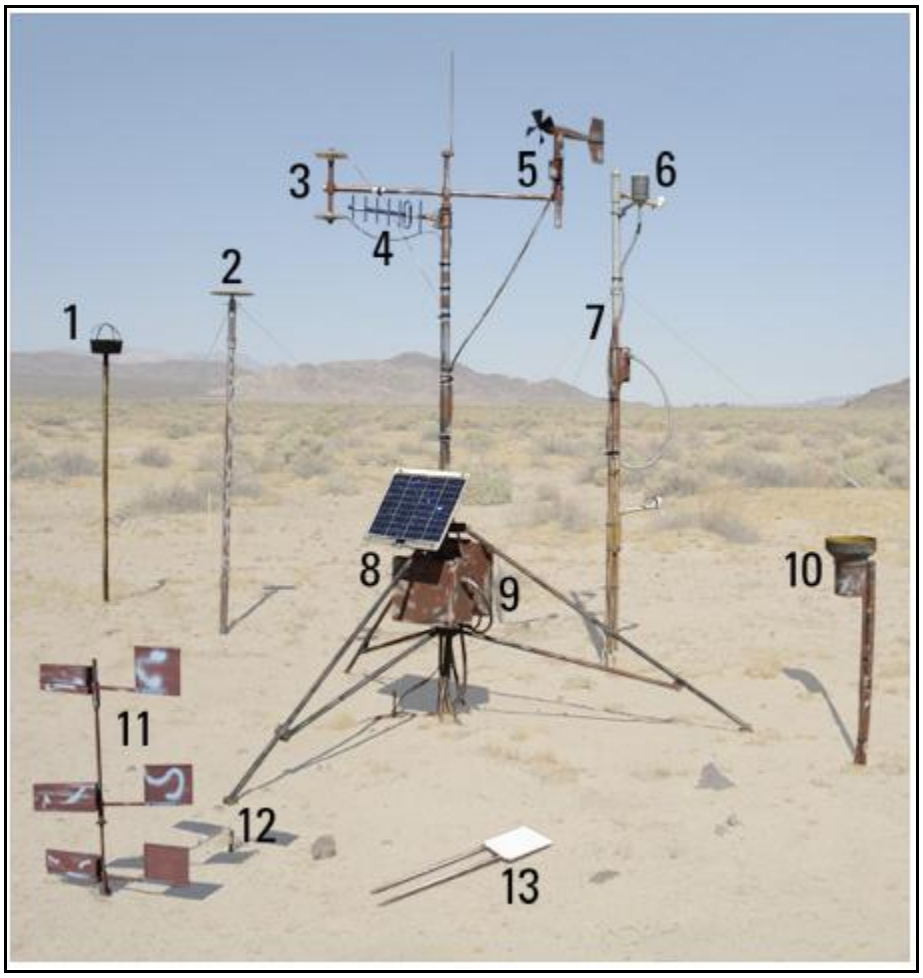

Figure 8. Equipment placed at the CLIMET stations [27].

Table 3. List of equipment employed at the CLIMET sites [27].

No.	Equipment name	Details
1	Marble Dust Trap	The design consists of a single-piece Teflon-coated angel-food cake pan painted black on the outside and mounted on a steel fence post about 2m above the ground.
2	Airfoil "Frisbee" Dust collector	It collects two types of samples: particles for analysis by microbeam methods and bulk deposition for bulk chemical analysis.
3	Pyranometer	Licor LI200X is used to measure solar radiation in the Light Spectrum Waveband 400 to 1100 nm.
4	Radio Telemetry system	The CR10X datalogger is connected to a radio transceiver and antenna.
5	Wind speed/direction	Campbell model 5103 is used to measure wind speed in the range of 0 to 60 m/s.
6	Air temperature and relative humidity	Campbell HMP35C is used to measure air temperature and relative humidity.
7	Temperature gradient sensor	Temperature gradient in degrees centigrade is measured between 1 and 3 meters above the ground surface.
8	Power module	It consists of a modular design that can contain one or two 26 amp gel-cell batteries charged by a 10 or 20 watt solar panel.
9	Datalogger	Campbell CR10X is used which has 12 single-ended-analog inputs, expandable with a multiplexer and 3 excitation outputs.
10	Rain gauge	Campbell TE525 Tipping bucket is used.
11	BSNE field dust sampler	This is a field dust sampler suitable for collecting airborne dust under natural field conditions. The sampler orients itself into erosive winds. Samples are taken at 15, 50, and 100 cm above the ground surface.
12	SENSIT erosion monitor	An experimental instrument (Sensit model H7) is used to detect particles in saltation.
13	Soil moisture	Campbell CS615 water content reflectometer is used.

The data for all sites, both Canyonlands and Mojave National Preserve, is available in 5-minute, hourly, daily and monthly datasets.

2.4. Summary

This chapter gives the details of data sources, equipment used, and locations. The NCCP and USGS are a few of the programs that generate and use huge amounts of data. The humidity data from NCCP and temperature data from USGS will be used later in this thesis. Thus, special methods need to be developed to transmit and store such data efficiently.

CHAPTER 3

DATA COMPRESSION

3.1. Introduction

Data compression is the science of representing information in a compact form by converting input into a compressed sequence of output symbols. Original input to these algorithms can be recovered using the corresponding decompression algorithm. The primary advantage of using compression methods is cheaper storage and transmission. The occurrence probability of equally sized input objects is non-uniform, which is exploited by data compression algorithms. Transmission costs can be considerably reduced by allocating shorter code words to probable objects and longer code words to less probable ones. Most contemporary systems transmit information using binary alphabets, 0 and 1 [28]. Data compression considerably reduces the number of bits used to store or transmit information by utilizing the fact that every data contains some level of redundancy, which can be removed or expressed differently, with the proviso of not altering its original meaning [29].

Recent advancements in communication networks has resulted in continuous transfer of copious amounts of data over communication links. Storage and communication cost of such massive data can be considerably reduced by employing compressions algorithms, consequently increasing the capacity of communication channels and allowing more data to be sent over a fixed bandwidth channel. Also, compressing a file increases capacity of the storage medium, making it feasible to store data at a higher rate [30].

Many data processing applications require storage of large volumes of data, and number of such applications is constantly increasing as the use of computers extends to new disciplines [30]. An aspect of data compression that should be examined is the exploding use of climate data, which requires efficient compression techniques. This thesis advocates better transmission and storage technologies for climate databases such as NCCP and USGS.

3.2. Codes and Communication

In the modern age, information is represented digitally and its transmission electronically. With the advent of Internet, large amount of data is exchanged globally. One way to represent information digitally is using Code. A code defines mapping between a set of input X and a set of sequences of symbols from a finite alphabet S. Codes are designed to serve the following purpose:

1. Encoding object X such that its output sequence S is tolerant to transmission errors and X can be reconstructed from 'S' in spite of it being altered during transmission. These are called error correcting codes.
2. Producing a compact representation of object X, and minimizing the length of output sequence. This is called data compression, and requires knowledge of probability distribution over input objects.
3. Protecting object X from interception. This is called encryption, and works by making decoding procedure depend on a secret without which decoding would either be impossible or computationally infeasible.

Under the work of Claude Shannon (1948), coding theory was developed which gave a formal way of measuring information and uncertainty. He laid the foundations for data compression and error correcting codes [28].

3.3. Basic concepts

Compression algorithms take an input X and generate a representation X_c that requires fewer bits compared to X. Conversely, decompression algorithms operate on compressed representation X_c to reconstruct X. There are different ways to evaluate a compression algorithm based on memory required, performance, speed, degree of compression and resemblance of reconstructed output to original input [31].

3.3.1. Compression Ratio (CR)

One way to measure efficiency of a compression algorithm is to calculate ratio of number of bits required to represent data before and after compression and is called the compression ratio.

Mathematically, it is defined as:

$$Compression\ Ratio = \frac{Uncompressed\ Size}{Compressed\ Size} \quad (1)$$

Percentage can be calculated using:

$$Compression\ \% = \left(1 - \left(\frac{Compressed\ \text{Size}}{Uncompressed\ \text{Size}}\right)\right) * 100 \quad (2)$$

For example, if a file is compressed to one-third its original size, it is said to be 67% compressed [31].

3.3.2. Types of compression

Based on the required degree of reconstruction, data compression algorithms can be classified into two broad categories:

- Lossless compression
- Lossy compression

Lossless ensures output Y is identical to input X and lossy in which Y is slightly different from X [31].

3.4. Lossless compression

In Lossless compression techniques, the entire input can be reconstructed from the compressed output, i.e., there is no change in information, neither during compression nor decompression. It is used for applications requiring exact representation of original data from compressed data. One of the major uses of lossless compression is for text compression, where small differences can result in significantly different meaning. Such compression

methods are used for storing database records, medical images, spreadsheets, text and images specifically preserved for legal purposes, and word file processing where loss of even a single bit is intolerable [31].

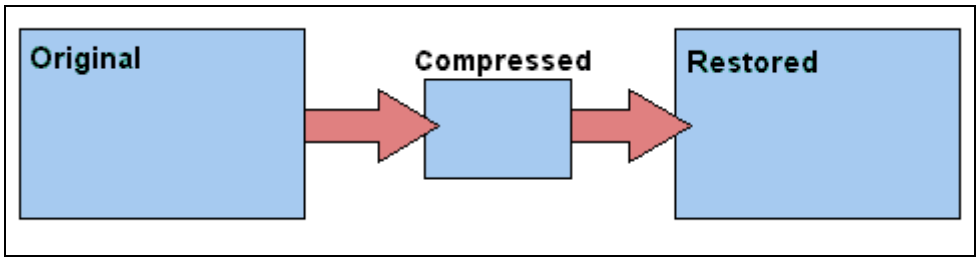

Figure 9. Lossless compression [32].

The basic working principle of lossless compression algorithms is the condensation of redundant information using statistical modeling techniques. These techniques calculate the probability of occurrence of a character or phrase in a file and assign shortest codes to the most recurring ones [33].

Till date, various lossless data compression algorithms have been proposed and used including Huffman Coding, Run Length Encoding, Arithmetic Encoding and Dictionary Based Encoding [34]. This thesis will use Huffman Coding for lossless compression of climate data.

3.5. Lossy compression

In Lossy compression techniques, original input cannot be completely reconstructed from the compressed output, i.e., there is loss of information during compression and/or decompression process. The upside for accepting this distorted reconstruction is much higher compression ratio. It is used in applications where exact reconstruction of output is not necessary like audio, video, images and detailed graphics for screen design (computers, TVs, projector screens). Lossy methods provide high degrees of compression resulting in smaller

compressed files with some amount of visual loss on restoration.

For example, a varying degree of loss in speech can be tolerated, depending on the quality requirements. If reconstructed speech is expected to be of similar quality as a telephonic conversation, significant losses can be tolerated. However, if reconstructed speech needs to be of the highest quality, tolerance is much lower. Similar compression scheme is used for video compression as long as the difference in output does not result in annoying artifacts [31]. Lossy compression algorithms take advantage of an inherent limitation of the human eye and discard information that cannot be detected by most users [33].

In lossy compression applications it is required to represent a large set of values with much smaller set, and this process is called quantization. This is the loss in lossy compression. The set of inputs for quantization process can be scalar or vectors. It they are scalars it is called scalar quantization whereas if they are vectors it is called vector quantization [31].

3.5.1. Scalar quantization

Quantization is a simple process and design of a quantizer has a huge impact on the amount of compression obtained. Quantizer consists of two mappings: an encoder mapping and a decoder mapping. Encoder divides the range of values generated by the source into a number of intervals. A distinct codeword is allocated to each interval. All source outputs that fall into a particular interval are encoded by the encoder using the codeword representing that interval. The encoding mapping is irreversible, as there can be many distinct values that can fall in any given interval. Corresponding to each codeword generated by the encoder, a reconstruction value is generated by the decoder. There is no way of knowing which value in the interval was actually generated by the source because the codeword represents the entire interval. Construction of intervals and selection of reconstruction values is part of the decoder design which affects the quality of reconstruction [31].

3.5.2. Vector quantization

In vector quantization, source output is grouped into blocks or vectors. Vector of source output forms the input to a vector quantizer. For both encoder and decoder of a vector quantizer, a set of say, L-dimensional vectors, called the codebook of vector quantizer are available. Vectors in this codebook, known as code-vectors, are selected to be representative of vectors generated from the source outputs. A binary index is assigned to each code-vector. In the encoder, input vector is compared to each code-vector in order to find the one closest to one another. The elements of this code-vector are the quantized values of the source output. In order to inform the decoder about the code-vector found to be closest to the input vector, a binary index of the code-vector is transmitted or stored. As the decoder has exactly the same codebook, it can retrieve the code-vector given its binary index. The encoder may have to perform a considerable amount of computation in order to find the closest reproduction vector to the vector of source outputs, but decoding consists of a table lookup. This makes vector quantization a very attractive encoding scheme for applications for which resources available for decoding are considerably less than the resources available for encoding [31].

3.6. Differential encoding

Differential encoding is useful in cases where signal output (x_n) is varying slowly, i.e., successive signal samples do not differ by much, but is not zero either. It is used in applications which have correlation from sample to sample. It makes data to be transmitted dependent not only on the current bit (or symbol), but also the previous one. This means both dynamic range and variance of the sequence of differences $d_n = x_n - x_{n-1}$ are significantly smaller than that of the source output sequence. For correlated sources, distribution of d_n is highly peaked at zero. Thus, encoding the difference from one sample to the next rather than encoding the actual sample value is useful. Techniques that transmit information by encoding differences are called differential encoding techniques [31]. This method is used in this thesis

which is then followed by Huffman coding.

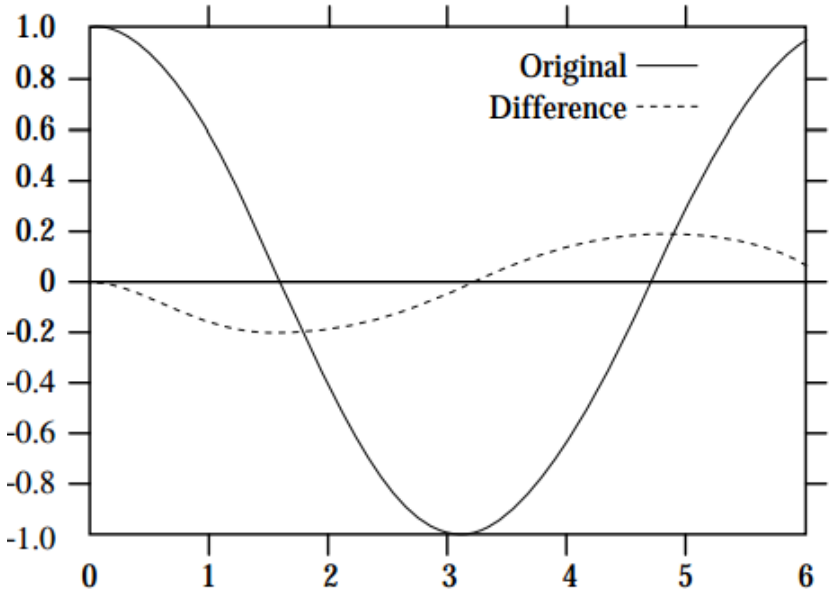

Figure 10. Sinusoid and sample to sample difference [31].

3.7. Huffman coding

Huffman coding is an example of variable-length encoding scheme developed by David Huffman. Codes generated using these techniques or procedures are called Huffman codes. It is a particular type of optimal prefix code commonly used for lossless data compression. Output from Huffman's algorithm can be viewed as a variable-length code table for encoding a source symbol. This table is derived from the frequency of occurrence or estimated probability of each possible value of the source symbol [35].

Huffman Encoding Algorithms use the probability distribution of an alphabet in the source to develop code words for symbols. In order to calculate the probability distribution, frequency distribution of all the source characters is calculated. Code words are assigned according to the probabilities. Symbols with higher probability of occurrence are assigned shorter code words and symbols with lower probability are assigned longer code words. For

this task, a binary tree is built according to their probabilities using symbols as leaves and path to those leaves as code words.

Static Huffman algorithms and Adaptive Huffman algorithms are two families of Huffman coding. In Static Huffman algorithms, frequencies are calculated first and a common tree is then generated for both compression and decompression processes. Details of this tree are saved or transferred with the compressed file. On the contrary, in Adaptive Huffman algorithms, the symbol probabilities are not known apriori and get calculated as the source symbols are processed and has a separate tree for compression and decompression [36].

3.7.1. Compression

Compression works by creating a binary tree of nodes which can either be leaves or internal nodes. Initially, all nodes are leaf nodes, which contain the symbol itself, weight (frequency of appearance) of the symbol and optionally, a link to the parent node which makes it easy to read the code in reverse order, starting from a leaf node. Internal nodes contain symbol weight, links to at most two child nodes and an optional link to the parent node. As a common convention, bit '0' represents traversing the left child and bit '1' represents traversing the right child.

The process begins with two leaf nodes with the smallest probability. A new parent node is constructed from these leaf nodes such that its probability is the sum of probabilities of its children. Once child nodes are merged into a single parent, they are no more considered, whereas the parent is then merged with the third lowest probability node to form a new parent. This process continues until we end up with one single node at the top, called the root of the tree [35].

3.7.2. Decompression

The process of decompression consists of translating a stream of prefix codes to

individual byte values by traversing the Huffman tree, one node at a time, as each bit is read from the input stream. However, the Huffman tree must be first reconstructed. One method is to simply prepend the entire Huffman tree, bit by bit, to the output stream. For example, assuming that 0 represents a parent node and 1 a leaf node, whenever the latter is encountered the tree building routine simply reads the next eight bits to determine the character value of that particular leaf. The process continues recursively until the last leaf node is reached; at that point, the Huffman tree will thus be faithfully reconstructed. Since the compressed data can include unused "trailing bits", decompressor must be able to determine when to stop producing output. This can be accomplished by either transmitting the length of the decompressed data along with the compression model or by defining a special code symbol to signify the end of input [35].

3.8. Summary

This chapter summarizes the need for data compression and types of compression methods. The two methods, Differential encoding and Huffman coding used with climate data in this thesis are also explained in this chapter.

CHAPTER 4

ARTIFICIAL NEURAL NETWORKS

4.1. Introduction

Artificial Neural Networks (ANNs), commonly known as neural networks, are powerful brain-inspired computational models. These models attempt to mirror capabilities of the human brain. In a general form, a neural network is a machine intended to model the ways in which brain performs a specific assignment; it is usually employed using electronic components or software simulations on a digital computer. These models are used to approximate functions that are unknown. A neural network is an enormous, parallel-distributed processor that has a natural inclination for storing experiential knowledge and making it accessible for utilization; it is inspired by the functioning of a human brain. Neural networks are also referred to as neuro computers, parallel-distributed processors, etc [37].

ANNs are powerful tools for modeling. They can learn and identify correlated patterns between input datasets and corresponding target values. Once trained, ANNs can be used to predict the outcome of new independent input data [38]. This feature is employed in various areas, such as computing, medicine, engineering, and many others. Neural networks have a remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques [39]. A neural net resembles a brain in two respects: the learning process is used by the network to acquire knowledge, and the synaptic weights (inter-neuron connection strengths) are used to store the knowledge [37].

ANNs are composed of a set of artificial neurons, which are also the processing units interconnected with other neurons. The interconnection between neurons present in various layers of the system is referred to as a network. The weights represent a connection between

the neurons [40].

Neural networks are interesting because of their potential use in prediction and classification problems [38]. The procedure used to perform the learning process is called a learning algorithm. In order to achieve good performance, a massive interconnection of simple cells, called neurons, is employed in neural networks.

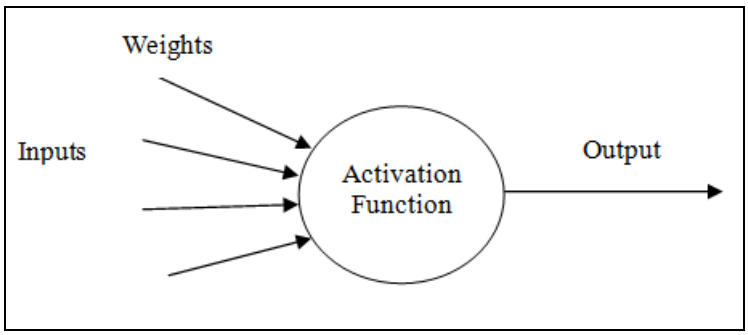

Figure 11. An artificial neuron [39].

4.2. Model of a neuron

A neural network consists of a set of connected cells: the neurons, which are information-processing units, fundamental to the operation of a neural network.

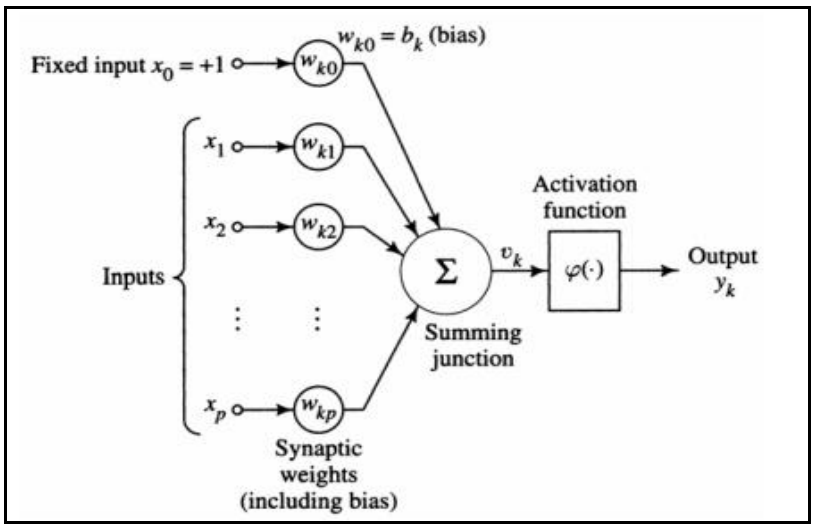

Figure 12. Non-linear model of a neuron [37].

There are three basic elements of a neuron model:

1. A set of synapses, each of which is characterized by a weight of its own. Specifically, a signal x_j at the input of synapse j connected to neuron k is multiplied by the synaptic weight w_{kj}. The first subscript refers to the neuron in question and the second subscript refers to the input end of the synapse to which the weight refers.

2. An adder for summing the input signals, weighted by the respective synapses of the neuron.

3. An activation function for limiting the amplitude of the output of a neuron. The model of a neuron also includes an externally applied bias (threshold) $w_{k0} = b_k$ that has the effect of lowering or increasing the net input of the activation function [37].

The neural networks are built from layers of neurons connected in such a way that one layer receives input from the preceding layer of neurons and passes the output on to the next layer. Neurons in the input layer receive the data and with the help of weighted links, the data is transferred to neurons in the first hidden layer. Data is mathematically processed and transfers the results to neurons in the next layer. Neurons in the last layer provide the network's output [39].

The j^{th} neuron in a hidden layer processes the incoming data (x_i) by calculating the weighted sum and adding a "bias" term (θ_j) according to:

$$net_j = \sum_{i=1}^{m} x_i * w_{ij} + \Theta_j \qquad j = (1, 2, 3, n) \qquad [41]$$

Neural networks are capable of adaptation to given data and are capable of generalization, even when the input data set contains noise or missing values.

4.3. Characteristics of neural networks

- They can map input patterns to their associated output patterns; that is, they exhibit mapping capabilities.

- They can predict new outcomes from past trends; that is, they possess the capability to generalize.
- They learn by example. They are trained with known examples of a problem before they are tested with unknown instances of the problem.
- They can process information in parallel, at high speeds, and in a distributed manner [38].

4.4. Learning process

A neural network has the ability to learn from its environment through an iterative process of adjustments applied to its synaptic weights and to improve its performance through learning. Learning methods in neural networks can be broadly classified into three basic types: supervised, unsupervised and reinforced. Each learning paradigm has many training algorithms.

4.4.1. Supervised learning

Supervised learning is a machine learning technique in which every input pattern used to train the network is associated with an output pattern, which is the target pattern. The training data is comprised of pairs of input and desired output values that are represented in data vectors. It is also referred to as classification, with a wide range of classifiers available; for example, multilayer perceptron, support vector machines, k-nearest neighbor algorithm, Gaussian mixture model, Gaussian, naive bayes, and decision trees. This process is carried out in various steps. The first step is to determine the type of training example. In the second step, a training dataset needs to be gathered. In the third step, the chosen training dataset needs to be described to the chosen neural network model. The fourth step is learning and after learning, performance of the learned neural network is tested with the test (validation) data set. The data, which was not introduced to the artificial neural network while learning, is

contained in the validation data set [39].

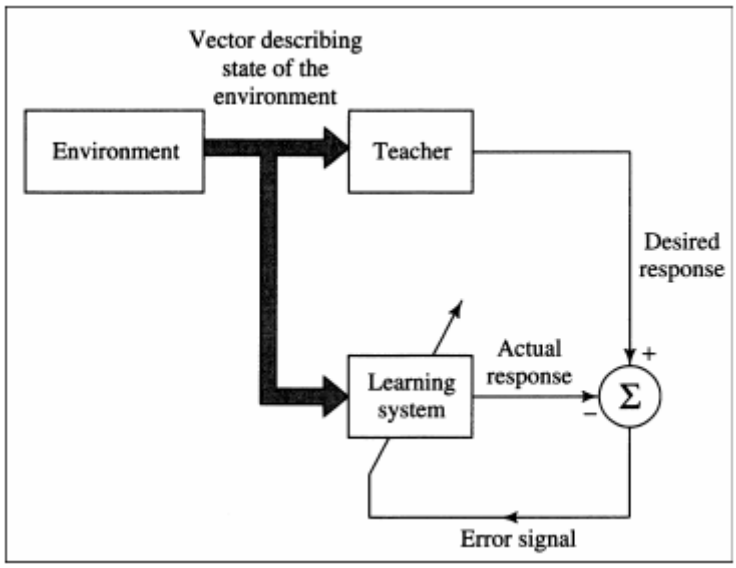

Figure 13. Supervised learning [37].

4.4.2. Unsupervised learning

Unsupervised learning is a machine learning technique in which the target output is not presented to the network. This learning process is used in applications which deal with estimation, such as statistical modeling, filtering, blind source separation and clustering. In unsupervised learning, we seek to determine how the data is organized. A provision is made for a task-independent measure of the quality of representation that the network is required to learn and the free parameters of the network are optimized with respect to that measure [37]. Self-organizing maps are the ones that commonly use unsupervised learning algorithms.

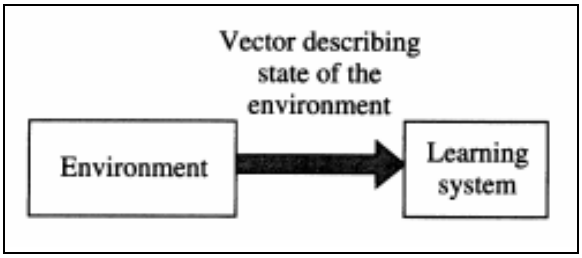

Figure 14. Unsupervised learning [37].

4.4.3. Reinforced learning

This type of learning is a machine learning technique that sets the parameters of an artificial neural network, where data is usually not given but generated by interactions with the environment. It is usually used as a part of an artificial neural network's overall learning algorithm. Reinforced learning is particularly suited for problems which include a long-term versus short-term reward tradeoffs. It has been applied successfully to various problems, including robot control, telecommunications, games like chess and other sequential decision making tasks [39].

4.5. Feed forward network

There are various types of ANNs, such as the feed forward neural network, Radial basis function network, Elman neural network, and Cascade feed forward neural network. The feed forward neural network is considered in this study. In this network, information flows from the input layer to the final output layer via zero or more hidden layers and the flows is unidirectional. There is no feedback loop, i.e., output of one layer does not affect that same or previous layer.

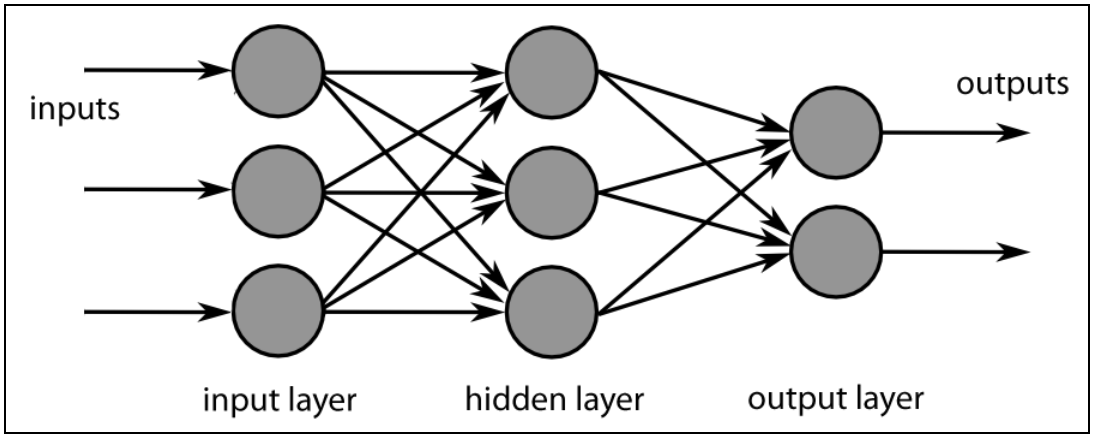

Figure 15. Feed forward neural network design [42].

4.6. Multi layer perceptron (MLP)

The most common neural network model is the multi layer perceptron. A graphical representation of an MLP is shown in Figure 16. This type of neural network is known as a supervised neural network because it requires a desired output in order to learn. The goal of MLP is to create a model that correctly maps the input to the output using historical data so that the model can then be used to produce the output when the desired output is unknown.

A set of external inputs feed the network. There are zero or more immediate hidden layers. The nodes in the input layer of the network supply respective elements of the activation pattern, which constitute the input signals applied to the neurons in the second layer (i.e., the first hidden layer). The output signals of the second layer are used as inputs to the third layer, and so on for the rest of the network. The set of output signals of the neurons in the output layer of the network constitutes the overall response of the network to the activation pattern supplied by the source nodes in the input layer [37]. MLP uses linear combination functions in the input layers and generally uses sigmoid activation functions in the hidden layers.

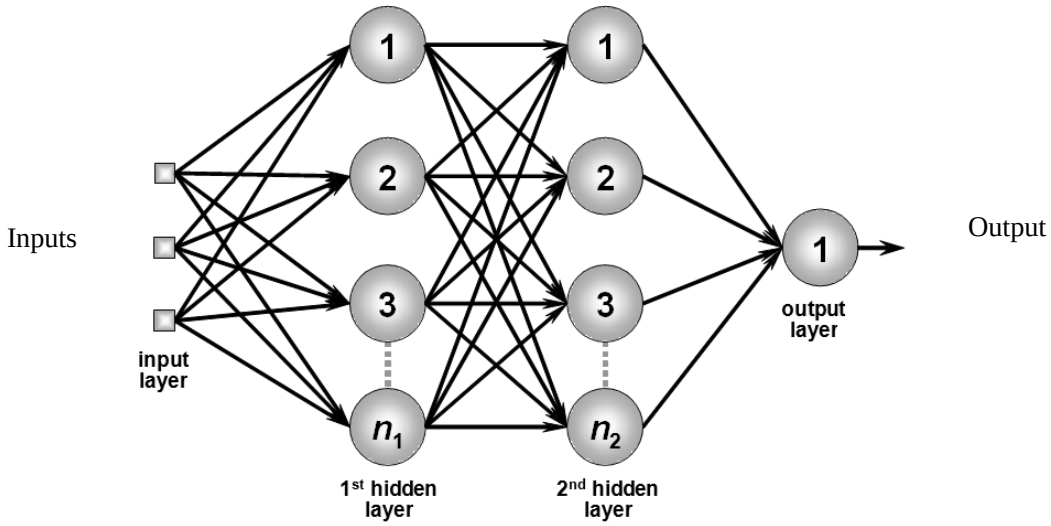

Figure 16. A fully connected multilayer feedforward network [43].

A neural network is said to be fully connected if every node in each layer of the network is connected to every other node in the adjacent forward layer. MLP utilizes back propagation algorithm, which is the standard algorithm for supervised learning pattern recognition process.

4.7. Back propagation algorithm

The task of back propagation is to train a multilayer perceptron (MLP), which is a loop-free network. This is used in conjunction with an optimization method such as gradient descent. This method calculates the gradient of a loss function with respect to all the weights in the network. The gradient is fed to the optimization method, which in turn, uses it to update the weights in an attempt to minimize the loss function.

Back propagation requires a known, desired output for each input value in order to calculate the loss function gradient. It is therefore usually considered a supervised learning method. It requires that the activation function used by the artificial neurons be differentiable. The goal of any supervised learning algorithm is to find a function that best maps a set of inputs to their correct outputs. The motivation for developing the back

propagation algorithm was to find a way to train a multi-layered neural network in such a way that it can learn the appropriate internal representations to allow it to then learn any arbitrary mapping of input to output [44].

With back propagation, the input data is repeatedly presented to the neural network. With each presentation, the output of the neural network is compared to the desired output and an error is computed. This error is then fed back (back propagated) to the neural network and used to adjust the weights such that the error decreases with each iteration; thus, the neural model gets closer and closer to producing the desired output. This process is known as "training" [45].

The back propagation learning algorithm can be divided into two phases: propagation and weight update.

4.7.1. Phase 1: Propagation

Each propagation phase involves following steps:

1. Forward propagation of a training pattern's input through the neural network in order to generate the propagation's output activations.

2. Backward propagation of the propagation's output activations through the neural network using the training pattern target in order to generate the deltas of all output and hidden neurons.

4.7.2. Phase 2: Weight update

For each weight-synapse, following steps are followed:

1. Multiply its output delta and input activation to get the gradient of the weight.

2. Subtract a ratio (percentage) of the gradient from the weight.

This ratio (percentage) influences speed and quality of learning; it is called the learning rate. The greater the ratio, the faster the neuron trains; the lower the ratio, the more accurate the training is. The sign of the gradient of a weight indicates where the error is increasing; this is

why weight must be updated in the opposite direction. Phase 1 and 2 are repeated until performance of the network is steady and satisfactory; this experiment defines "satisfactory" as when the occurrence of error is minimal [44].

4.8. Benefits of neural networks

A neural network derives its computing power through, first, its massively parallel distributed structure and, second, its ability to learn and generalize. The neural network's capability to produce rational outputs for inputs not encountered during training (learning) is termed "generalization." This feature of neural networks enables them to solve complex problems that are currently intractable. Use of neural networks offers following useful properties and capabilities:

1. *Nonlinearity*: A neuron is a nonlinear device, thus, a neural network, which is made up of an interconnection of neurons, is itself nonlinear.

2. *Input-output mapping*: Supervised learning involves modification of the synaptic weights of a neural network by applying a set of training samples. Each sample consists of a unique input signal and the corresponding desired response. The network is presented with a sample picked at random from the set, and the synaptic weights of the network are modified so as to minimize the difference between the desired response and the actual response. The training of the network is repeated for many samples in the set until the network reaches a steady state where there are no further significant changes in the synaptic weights. The previously applied training samples may be reapplied during the training session, usually in a different order. Thus, the network learns from the samples by constructing input-output maps for the problems at hand.

3. *Contextual information*: Knowledge is represented by the structure and activation state of a neural network. Every neuron in the network is potentially affected by the global activity of all other neurons in the network. Consequently, contextual information is dealt naturally by a

neural network [37].

4.9. Summary

In this chapter, ANNs have been described. The type of neural network model used in this study, i.e., the feed forward neural network, has also been described. The algorithm used for this study has also been covered in this chapter.

CHAPTER 5

DATA COMPRESSION USING ARTIFICIAL NEURAL NETWORKS

5.1. Overview

This chapter deals with results of compression. Section 5.2 states the parameters studied; Section 5.3 describes sites considered for the study; Section 5.4 describes the ANN model; Section 5.5 proposes a new method and Section 5.6 presents the results.

5.2. Parameters studied

The following climate data parameters are studied in this thesis:

- Humidity
- Temperature

Humidity data is used from 'Nevada Climate Change Data' and temperature data is used from 'United States Geological Survey' portal. These datasets are then used with Artificial Neural Networks.

5.2.1. Humidity dataset

Humidity data captured for 2013 and 2014 is downloaded from NCCP. Input file has daily readings for nine parameters namely, temperature, pressure, incoming shortwave, incoming long wave, outgoing shortwave, outgoing long wave, solar radiation, time in minutes and month, whereas the output file has readings for humidity. Both input and output files consist of time series data for each year. Data from 2013 is used for training the ANN model and data from 2014 is used for simulation. Simulation input files have 2014 data for 1 and 10 minute intervals. Each experiment yields humidity data for 2014 and is measured in percentage (%).

Table 4. Observed parameters and units of measurement for humidity model.

PARAMETERS OBSERVED	UNITS OF MEASUREMENT
Humidity	%
Temperature	Deg C
Pressure	Pa
Solar Radiation	W/m^2
Incoming shortwave and long wave	W/m^2
Outgoing shortwave and long wave	W/m^2

5.2.2. Temperature dataset

For this study, hourly readings of parameters from 1999 to 2012 is downloaded from USGS. Input file has eleven parameters namely year, day of year, hour, wind speed, wind direction, total particle count, soil moisture, peak wind speed, relative humidity, ground temperature, and rainfall. Output file has records for air temperature. Input and output is time series data. Data from 1999 to 2011 is used for training and data from 2012 is used for simulation. The unit of measurement is degree C (deg C).

Table 5. Observed parameters and their units of measurement for temperature model.

PARAMETERS OBSERVED	UNITS OF MEASUREMENT
Wind speed	m/s
Wind direction	Degrees
Soil moisture	%
Peak wind speed	m/s
Relative humidity	%
Ground temperature	Deg C
Rainfall	mm/hour

5.3. Sites considered

Study of humidity is performed at the following sites:

5.3.1. Snake Range West Subalpine

This site is part of the Snake Range of Nevada. Details of location are given in Table 6. For this location, two experiments were carried out, one with minute interval data and

another with 10 minute interval data.

Table 6. Details of Snake Range West Subalpine [19].

SITE NAME	LATITUDE (°)	LONGITUDE (°)	ALTITUDE (m)
Snake Range West Subalpine	38.90611	-114.30891	3353.4096

5.3.2. Sheep Range Mojave Desert Shrub

This site is part of the Sheep Range of Nevada. Details of the location are given in Table 7. For this location also, two experiments were carried out, one with minute interval data and another with 10 minute interval data.

Table 7. Details of Sheep Range Mojave Desert Shrub [19].

SITE NAME	LATITUDE (°)	LONGITUDE (°)	ALTITUDE (m)
Sheep Range Mojave Desert Shrub	36.4353453661066	-115.355850373052	893.064

The study of temperature is performed at the following sites:

5.3.3. North Soda Lake

This site is part of the Mojave National Preserve, California and its details are given in Table 8. Hourly data is used for this experiment.

Table 8. Details of North Soda Lake [27].

SITE NAME	LATITUDE (°)	LONGITUDE (°)	ALTITUDE (m)
North Soda Lake, Mojave National Preserve.	35° 13.479'	116° 04.125'	282 m

5.4. Artificial Neural Network (ANN) Model

As part of this study, an ANN model was developed based on past observations of

several meteorological parameters provided as input for training the model. Appropriate inputs were given to humidity and temperature models to calculate results via two experiments.

In the first experiment, effectiveness of multilayer perceptron ANN model was studied for the prediction of humidity data. The difference between the predicted values from the trained ANN model and original raw values from the data portal was calculated. Differential encoding followed by Huffman coding was then performed on this data to evaluate the compression ratio. Same process was followed for temperature data. In the second experiment, differential encoding followed by Huffman coding was directly performed on data extracted from the data portals. This experiment was also carried out on humidity and temperature data. Finally, results from both experiments were compared and analyzed.

5.4.1. Humidity model

Humidity data for this study was obtained from NCCP for years 2013 and 2014, and for locations Snake Range West Subalpine and Sheep Range Mojave Desert Shrub. For each of these locations, two experiments were conducted, one for a minute interval data and another for 10 minute interval data.

ANN model in this experiment used multilayer perceptron model comprising of hidden layers and neurons, though number of layers and neurons varied in each experiment. Input and output parameters are given in Table 9. Input data from 2013 was used for training the ANN model and data from 2014 for simulation.

Table 9. Input and output parameters for humidity model.

Input to ANN	Time in minutes, Month, Temperature, Pressure, Incoming Shortwave, Incoming Long Wave, Outgoing Shortwave, Outgoing Long Wave, Solar Radiation.
Output	Humidity

5.4.1.1. Model for Snake Range West Subalpine

ANN model was trained with humidity data for the year 2013. For the experiment where 1 minute interval dataset was used, model uses 1 hidden layer and 81 neurons as shown in Figure 17, whereas for the 10-minute interval dataset, model uses 3 hidden layers and 18 neurons as shown in Figure 18.

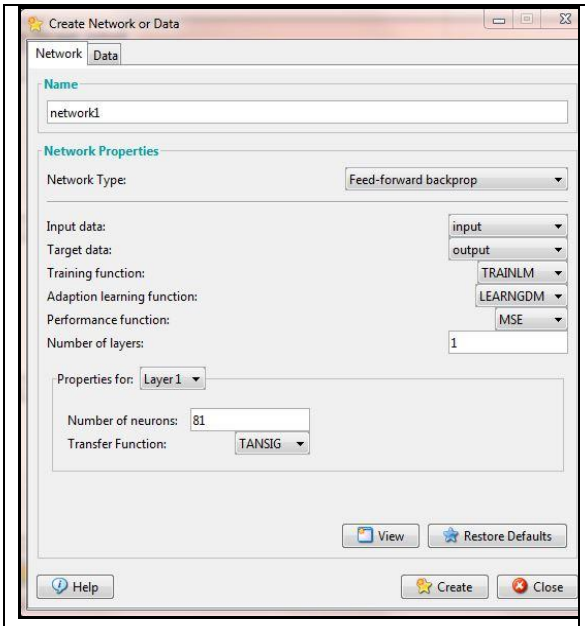

Figure 17. ANN Network model for 1-minute data for Snake Range West Subalpine.

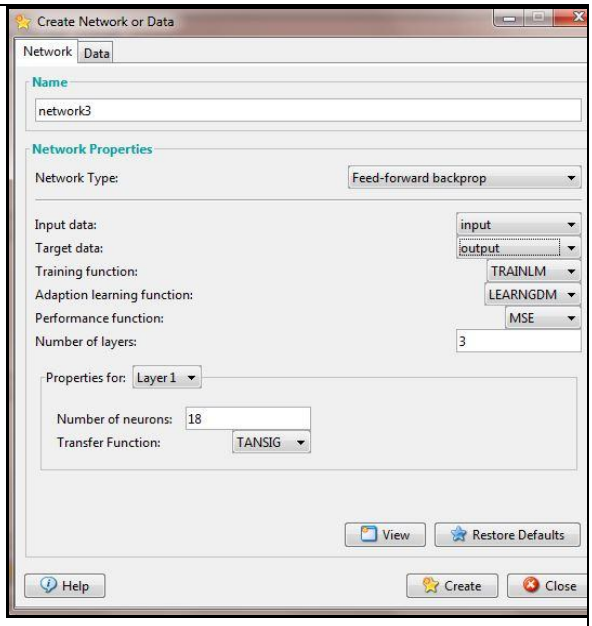

Figure 18. ANN Network model for 10-minute data for Snake Range West Subalpine.

5.4.1.2. Model for Sheep Range Mojave Desert Shrub

ANN model for this location was trained with humidity data for 2013. For the sheep range model, experiment where 1 minute interval dataset is used, model uses 3 hidden layers

and 47 neurons as shown in Figure 19, whereas for the 10 minute interval dataset, the model uses 2 hidden layers and 22 neurons as shown in Figure 20.

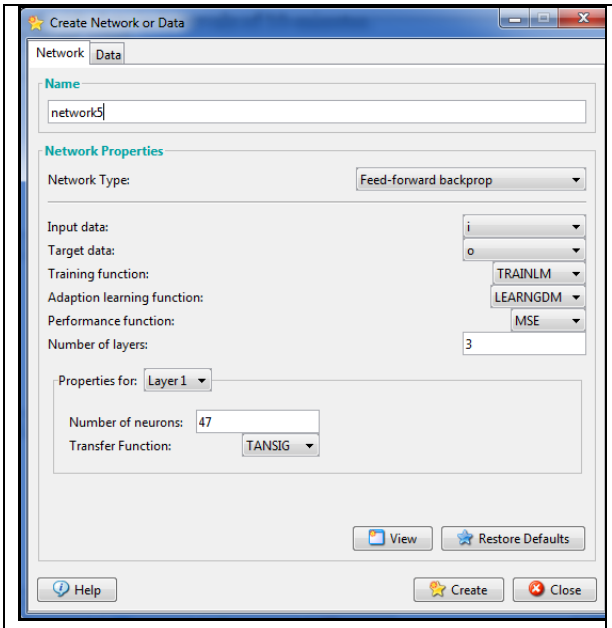

Figure 19. ANN Network model for 1-minute data for Sheep Range Mojave Desert Shrub.

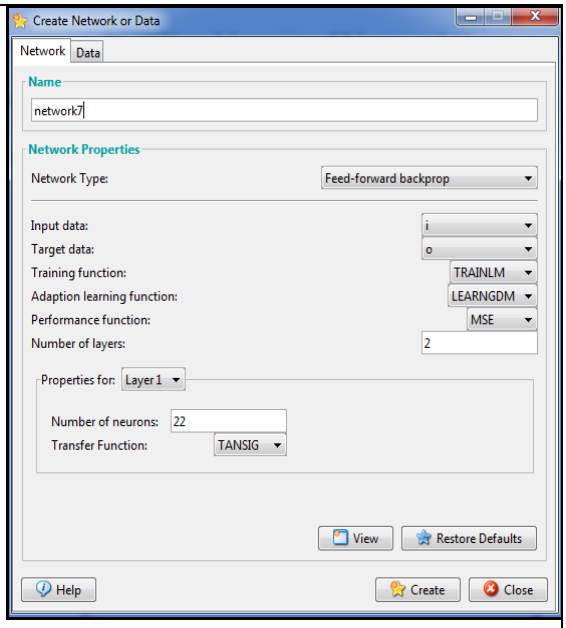

Figure 20. ANN Network model for 10-minute data for Sheep Range Mojave Desert Shrub.

5.4.2. Temperature model

Temperature data from USGS was observed from 1999 to 2012 for the North Soda Lake location, which is part of Mojave National Preserve, California. This dataset consists of hourly data.

The multilayer perceptron model of ANN was used in this experiment. Input and output parameters to ANN are given in Table 10. In this case, the model comprised of 3 hidden layers and 33 neurons as shown in Figure 21. Input data from 1999 to 2011 was used for training the model and data from 2012 for simulation.

Table 10. Input and output parameters for temperature model.

Input to ANN	Year, Day of year, hour, Wind speed, Wind direction, Total particle count, Soil moisture, Peak wind speed, Relative humidity, Ground temperature, Rainfall.
Output	Air temperature

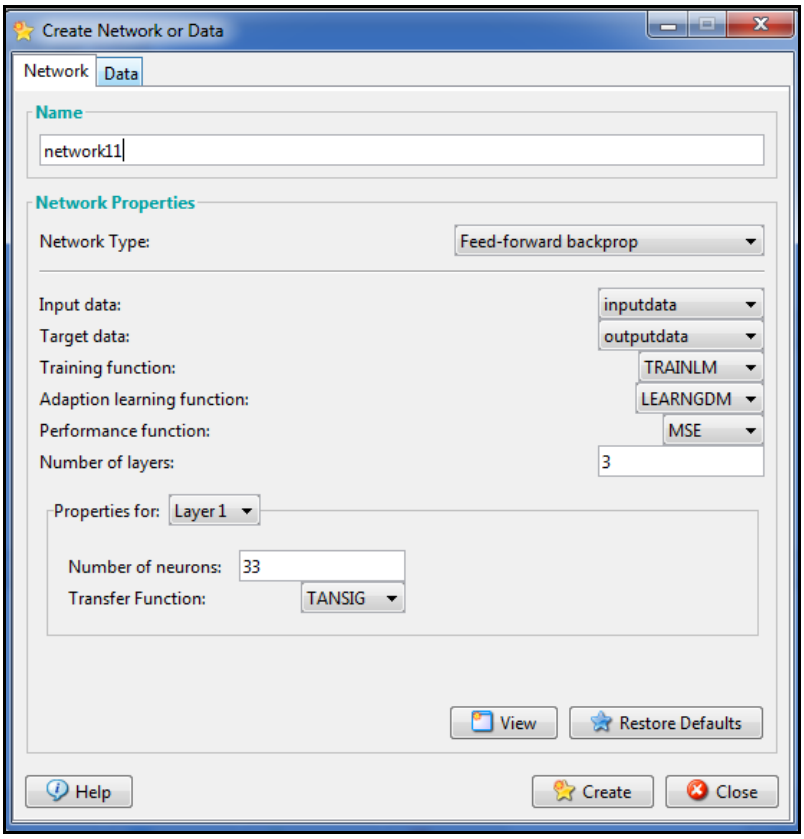

Figure 21. ANN model for temperature data.

5.5. Proposed method

5.5.1. Method 1

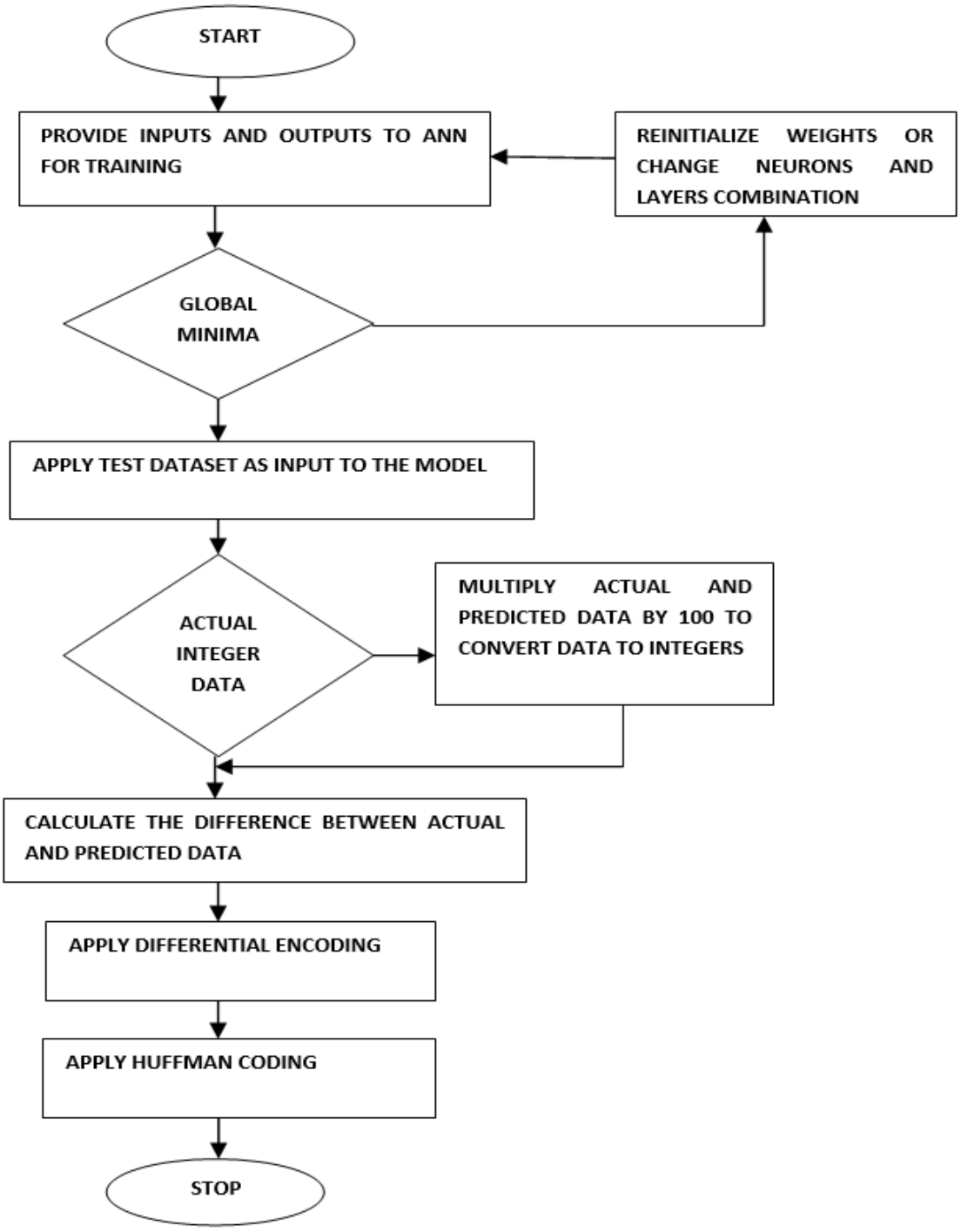

Figure 22. Flowchart for compression algorithm.

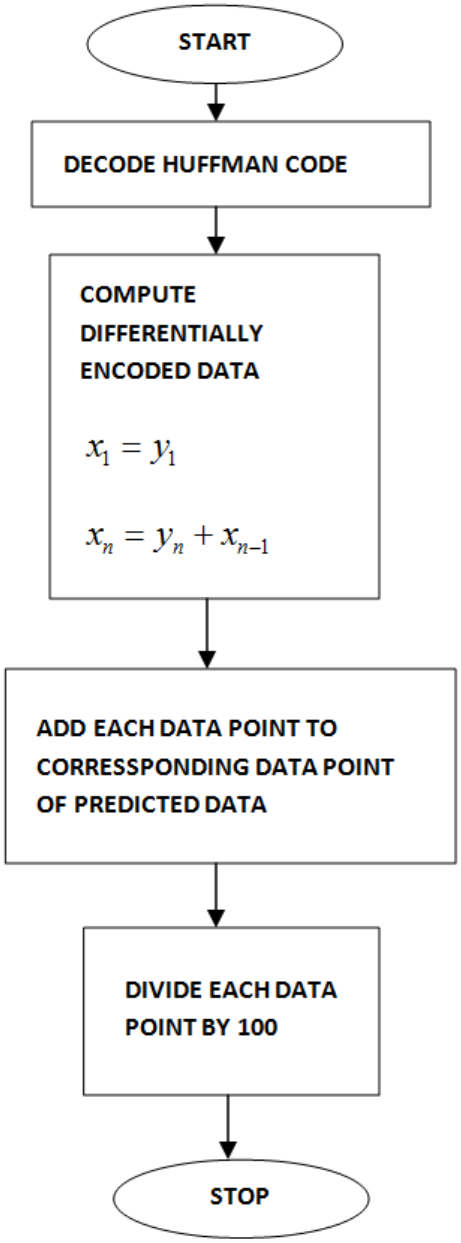

Figure 23. Flowchart for decompression algorithm.

5.5.1.1. Parameter 1: Humidity

Input and output for 2013 is presented to the multi-layer perceptron model of ANN for training. After training, monthly input data for 2014 is provided to the trained model to obtain predicted values of humidity for 2014. After multiple iterations, best performance (mean square error) is noted and corresponding predicted values of humidity are saved. Humidity is

represented as floating point data with two digits after the decimal point and as part of pre-processing, actual and predicted values are multiplied by 100 to convert them to integers. Difference between actual and predicted humidity data is then calculated. Next step is to use differential encoding where sample to sample difference is calculated using the formula:

$$y_1 = x_1, \tag{3}$$

$$y_n = x_n - x_{n-1} \; ; \; n = 2,3,4.... \tag{4}$$

where "x" is the source output and "y" is the sequence of differences of the source output. When the difference is small, compression is observed to improve. Finally, Huffman coding is applied to encode the data using fewer bits, thus giving lossless compression.

For decompression, first step is to decode the Huffman code. Next, the differentially encoded data is retrieved using:

$$r_1 = s_1 \; , \tag{5}$$

$$r_n = s_n + r_{n-1} \; ; \; n = 2,3,4.... \tag{6}$$

where "r" is differentially encoded data and "s" is data obtained after huffman code is decoded.

Each data point is added to corressponding data point of predicted data. This process is continued until the last data point is encountered. Lastly, every data point is divided by 100 to obtain the actual values. Thus, lossless compression and decompression is applied for humidity data.

For the first experiment, where 1 minute interval data is used, input file for a year is a 2 dimensional matrix of 9 x 525600 with a total of 4730400 data points, and output is a 1 dimensional data set with 44640 data points. For the second experiment, where the 10 minute interval data values is used, input file for a year is a 2 dimensional matrix of 9 x 52560 with a total of 473040 data points, and output is a 1 dimensional data set with 4464 data points.

5.5.1.2. Parameter 2: Temperature

Hourly input and output from 1999 to 2011 is presented to the multilayer perceptron model of ANN for training. Input data from 2012 is fed to the trained model to provide the predicted temperature values for 2012. Similar to humidity data, temperature is also represented as floating point data with two digits after the decimal point and is multiplied by 100 to convert it to integer. Next, the difference between actual and predicted temperature data is calculated and differential encoding followed by Huffman coding is applied to encode the data using fewer bits.

For decompression, same procedure is used as in previous case. Thus, lossless compression and decompression for temperature data is obtained.

For this experiment, input is a 2 dimensional matrix of 11 x 113952 with a total of 1253472 data points, and output is a 1 dimensional data set with 8784 data points.

5.5.2. Method 2

Parameters used in method 1 i.e., humidity and temperature are used here too. In this case, instead of using predicted data, original values are used. The conventional method of differential encoding is used where sample to sample difference is calculated using the formula stated in method 1. It is followed by Huffman coding to evaluate the compression ratio. Results are calculated for both humidity and temperature datasets.

The results obtained from method 1 and 2 are then compared and analyzed.

5.6. Results

This section presents the results of experiments performed. Section 5.6.1 will give the results of humidity dataset and Section 5.6.2 presents the results of temperature dataset. Performance metrics used for this study are compression ratio (CR) and root mean square error (RMSE).

5.6.1. Results of humidity data

5.6.1.1. Sheep Range Mojave Desert Shrub

Table 11. MSE and RMSE vales for Sheep Range Mojave Desert Shrub.

Performance level	Sheep Range Mojave Desert Shrub	
	1- minute	10-minute
Mean Square Error	27.14	53.7
Root Mean Square Error	5.21	7.32

The plot of actual and predicted humidity data for the location Sheep Range Mojave Desert Shrub is given in Figure 24. It can be seen from the graph that the predicted value is closer to the actual value for each month which provides a good compression.

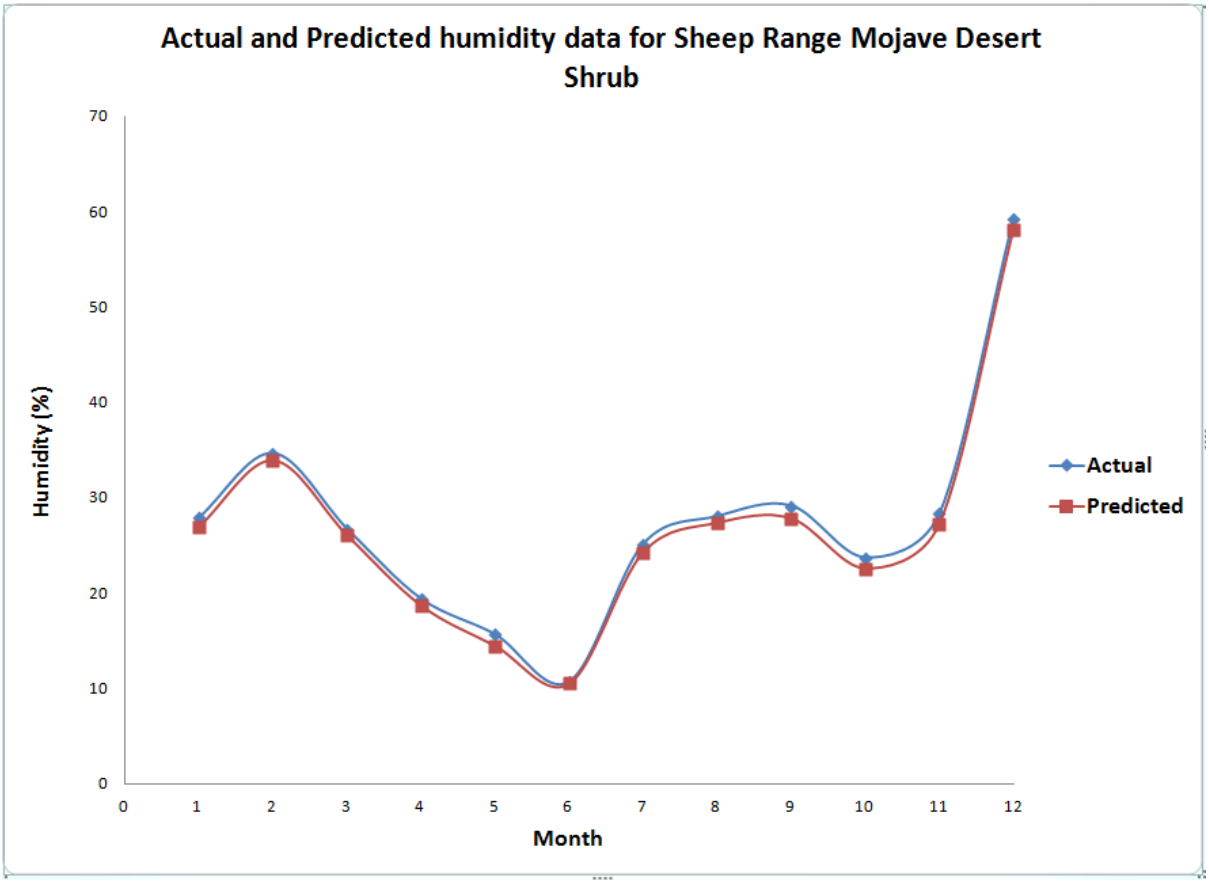

Figure 24. Actual and predicted humidity values for Sheep Range Mojave Desert Shrub.

Table 12. Monthly compression ratios for the year 2014 for minute interval humidity dataset for Sheep Range Mojave Desert Shrub.

MONTH	UNCOMPRESSED BITS	COMPRESSED BITS (using method 1)	COMPRESSION RATIO (using method 1)	COMPRESSED BITS (using method 2)	COMPRESSION RATIO (using method 2)
JANUARY	758880	159429	4.76	308117	2.46
FEBRUARY	685440	152659	4.49	288727	2.37
MARCH	758880	170152	4.46	323963	2.34
APRIL	734400	140421	5.23	305978	2.40
MAY	758880	134078	5.66	330605	2.29
JUNE	**758880**	**123596**	**6.14**	**301630**	**2.66**
JULY	803520	157245	5.11	319579	2.51
AUGUST	803520	163317	4.92	322520	2.49
SEPTEMBER	777600	167586	4.64	313694	2.48
OCTOBER	803520	151322	5.31	313227	2.42
NOVEMBER	777600	189197	4.11	295598	2.63
DECEMBER	803520	168453	4.77	304422	2.64

Figure 25 shows the comparison of obtained compression ratios using proposed method and conventional method for minute-interval dataset. It is seen from the graph that the monthly compression ratios using proposed method are greater than those obtained by conventional method. The maximum compression ratio using proposed method is 6.14 whereas using conventional method it is 2.66.

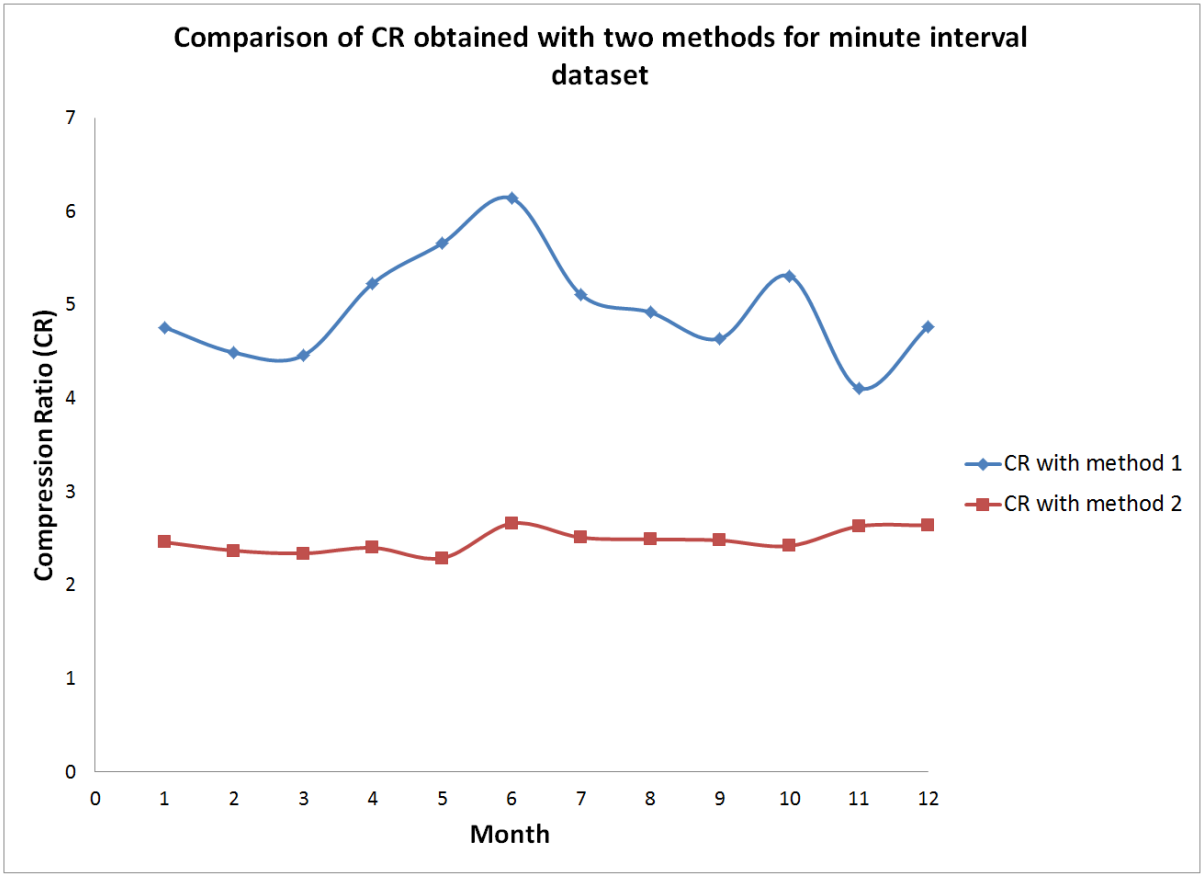

Figure 25. Comparison of compression ratio obtained with two methods for minute interval dataset of Sheep Range Mojave Desert Shrub.

Table 13. Monthly compression ratios for the year 2014 for 10-minute humidity for Sheep Range Mojave Desert Shrub

MONTH	UNCOMPRESSED BITS	COMPRESSED BITS (using method 1)	COMPRESSION RATIO (using method 1)	COMPRESSED BITS (using method 2)	COMPRESSION RATIO (using method 2)
JANUARY	75888	19164	3.96	39666	1.91
FEBRUARY	68544	17486	3.92	36056	1.90
MARCH	75888	18924	4.01	39515	1.92
APRIL	73440	16805	4.37	37189	1.97
MAY	75888	16426	4.62	39512	1.92
JUNE	**75888**	**14427**	**5.26**	**33878**	**2.24**
JULY	80352	18303	4.39	38629	2.08
AUGUST	80352	18557	4.33	38858	2.06
SEPTEMBER	77760	18692	4.16	38161	2.03
OCTOBER	80352	18514	4.34	38321	2.09
NOVEMBER	77760	18296	4.25	37550	2.07
DECEMBER	80352	19889	4.04	39018	2.05

Figure 26 shows the comparison of obtained compression ratios using proposed method and conventional method for 10-minute interval dataset. The maximum compression ratio using proposed method is 5.26 whereas using conventional method it is 2.24.

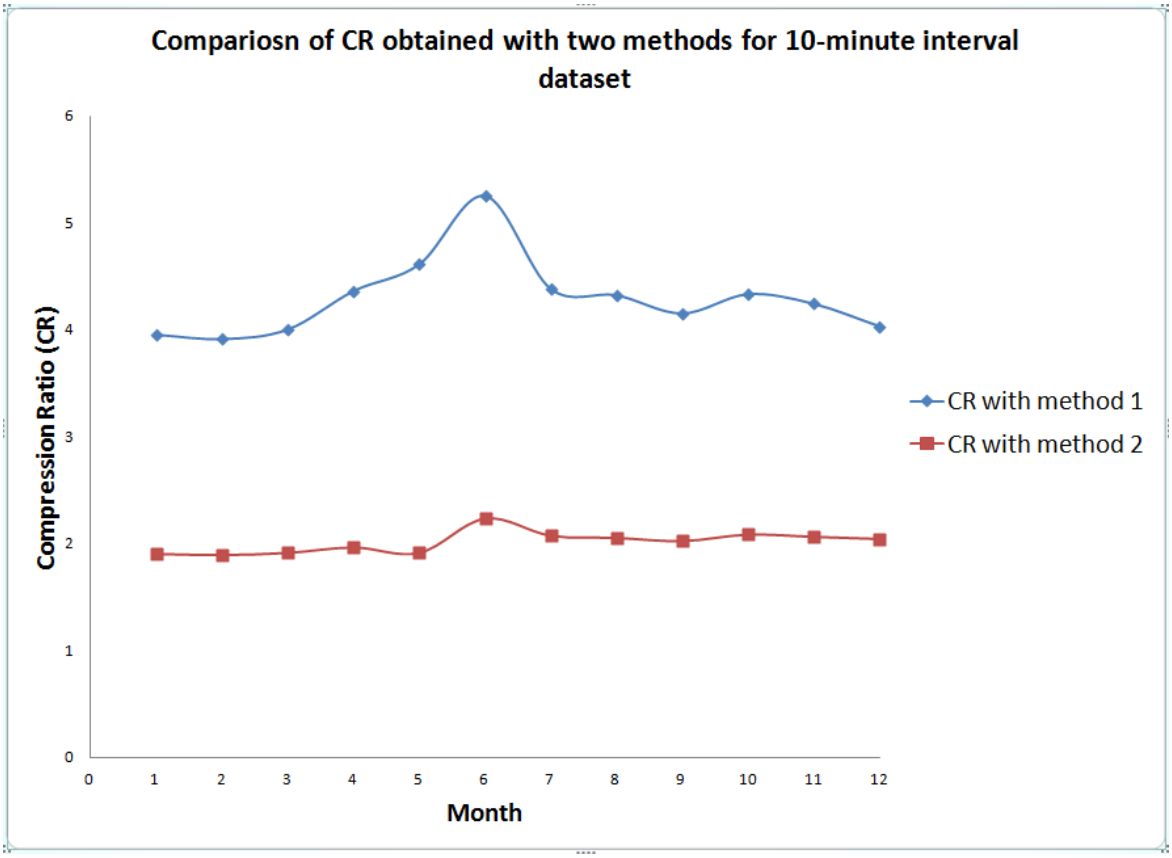

Figure 26. Comparison of compression ratio obtained with two methods for 10-minute interval dataset of Sheep Range Mojave Desert Shrub.

5.6.1.2. Snake Range West Subalpine

Table 14. MSE and RMSE values for Snake Range West Subalpine.

Performance level	Snake Range West Subalpine	
	1- minute	10-minute
Mean Square Error	42.5	90.5
Root Mean Square Error	6.52	9.5

The plot of actual and predicted humidity data for the location Snake Range West Subalpine is given in Figure 27. It can be seen from the graph that the predicted value is close

to the actual value for each month which provides a good compression. Predicted value is closest to the actual value for the month of June for which the maximum compression ratio is obtained.

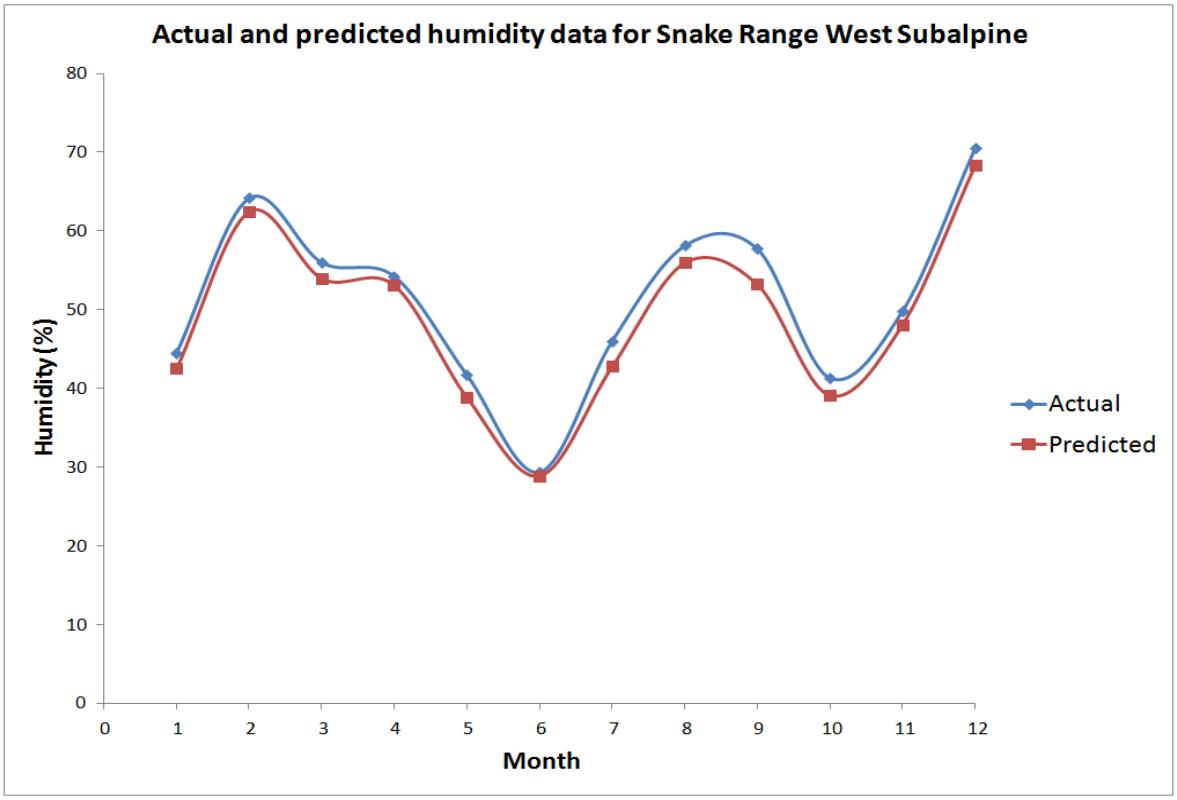

Figure 27. Actual and predicted humidity values for minute interval dataset for Snake Range West Subalpine.

Table 15. Monthly compression ratios for the year 2014 for 1-minute humidity data for Snake Range West Subalpine.

MONTH	UNCOMPRESSED BITS	COMPRESSED BITS (using method 1)	COMPRESSION RATIO (using method 1)	COMPRESSED BITS (using method 2)	COMPRESSION RATIO (using method 2)
JANUARY	758880	176074	4.31	366769	2.07
FEBRUARY	685440	149659	4.58	293386	2.33
MARCH	758880	179404	4.23	364289	2.08
APRIL	734400	144566	5.08	275056	2.67
MAY	758880	183304	4.14	387163	1.96
JUNE	**758880**	**129723**	**5.85**	**274956**	**2.76**
JULY	803520	144258	5.57	370556	2.16
AUGUST	803520	179357	4.48	363373	2.21
SEPTEMBER	777600	167948	4.63	342671	2.26
OCTOBER	803520	184293	4.36	371492	2.16
NOVEMBER	777600	164397	4.73	338991	2.29
DECEMBER	803520	194086	4.14	380815	2.11

The comparison of compression ratios obtained using proposed method and conventional method for minute-interval dataset is shown in Figure 28. Monthly compression ratios using proposed method are greater than those obtained by conventional method. The maximum compression ratio using proposed method is 5.85 whereas using conventional method it is 2.76.

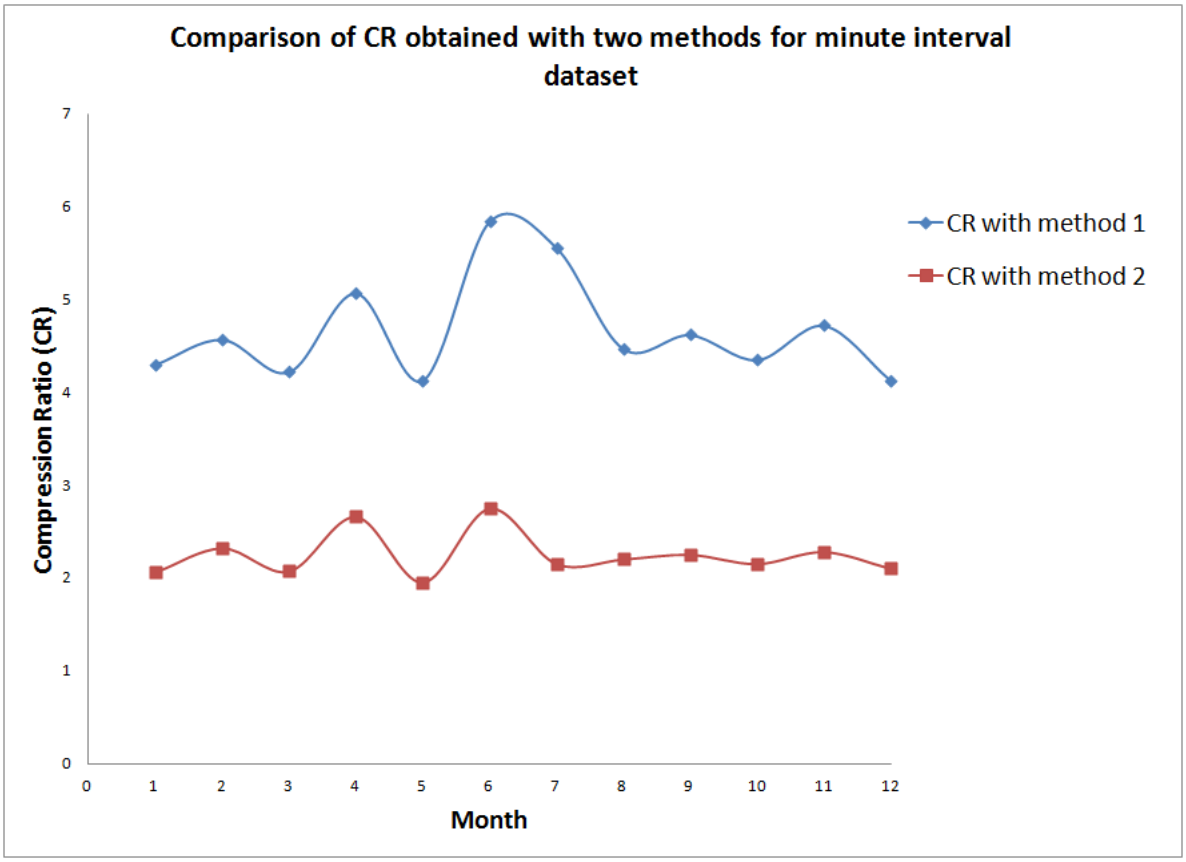

Figure 28. Comparison of compression ratio obtained with two methods for minute interval dataset of Snake Range West Subalpine.

Table 16. Monthly compression ratios for the year 2014 for 10-minute humidity for Snake Range West Subalpine.

MONTH	UNCOMPRESSED BITS	COMPRESSED BITS (using method 1)	COMPRESSION RATIO (using method 1)	COMPRESSED BITS (using method 2)	COMPRESSION RATIO (using method 2)
JANUARY	75888	26627	2.85	39449	1.92
FEBRUARY	68544	22327	3.07	31935	2.15
MARCH	75888	26815	2.83	38698	1.96
APRIL	73440	26901	2.73	39449	1.86
MAY	75888	26721	2.84	40351	1.88
JUNE	**75888**	**19558**	**3.88**	**33578**	**2.26**
JULY	80352	27423	2.93	40995	1.96
AUGUST	80352	26606	3.02	39005	2.06
SEPTEMBER	77760	25663	3.03	37931	2.05
OCTOBER	80352	27423	2.93	41418	1.94
NOVEMBER	77760	25578	3.04	37565	2.07
DECEMBER	80352	24572	3.27	38817	2.07

The comparison of compression ratios obtained using proposed method and conventional method for 10-minute interval dataset is shown in Figure 29. The maximum compression ratio using proposed method is 3.88 whereas using conventional method it is 2.26.

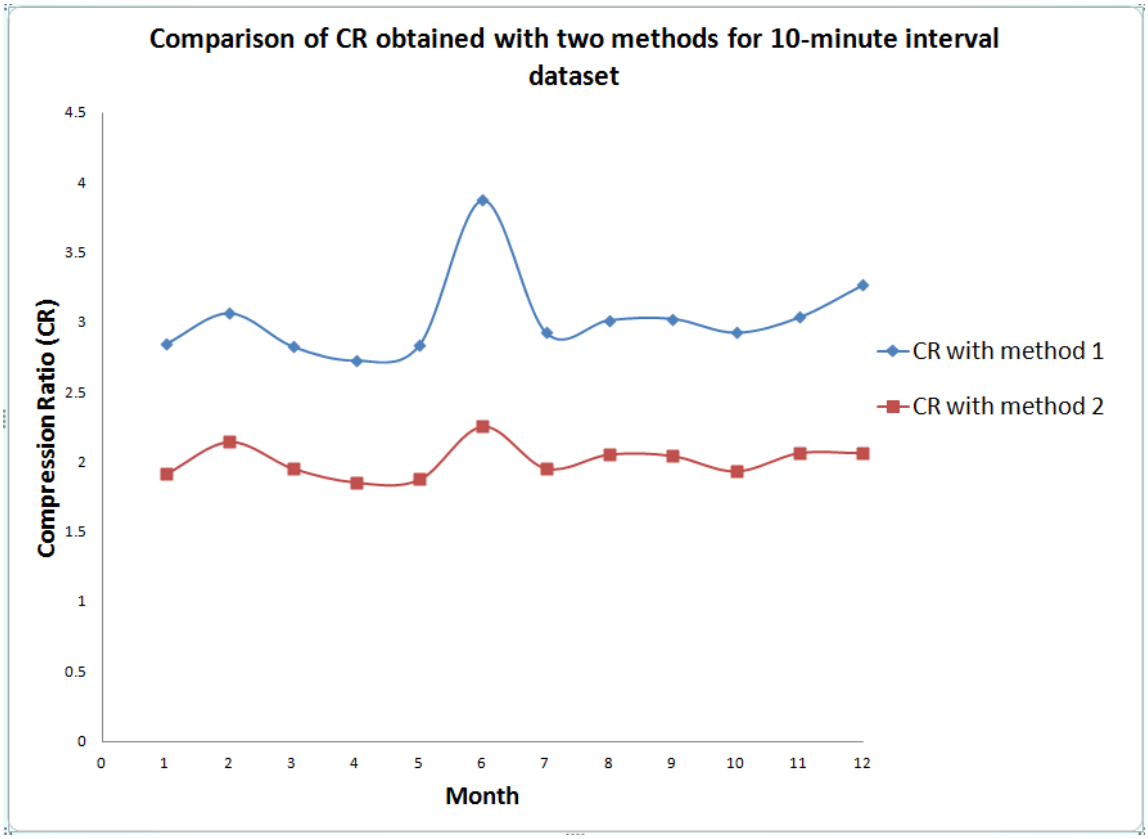

Figure 29. Comparison of compression ratio obtained with two methods for 10-minute interval dataset of Snake Range West Subalpine.

5.6.2. Results of temperature data

5.6.2.1. North Soda Lake

The plot of actual and predicted temperature data for the location North Soda Lake is given in Figure 30. It can be seen from the graph that the predicted value is closer to the actual value for each month which provides a good compression.

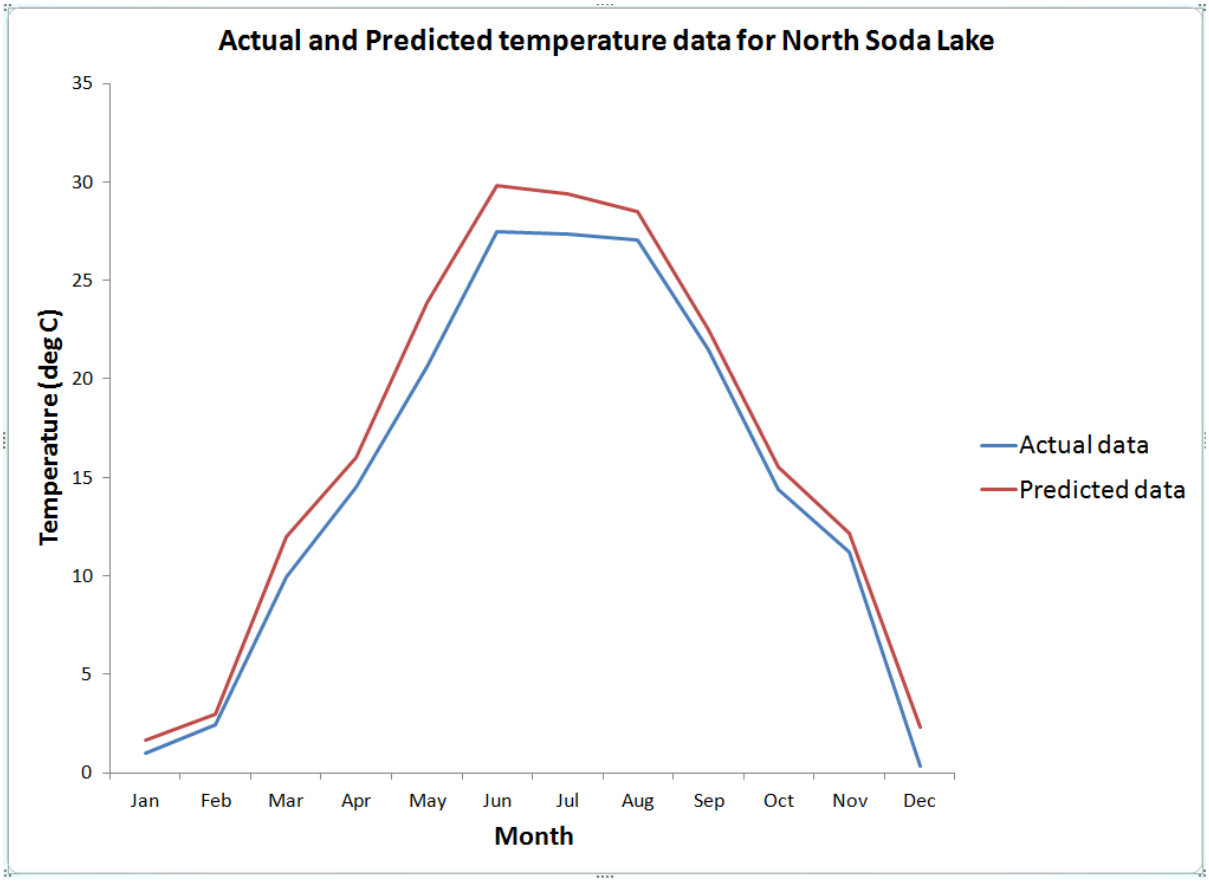

Figure 30. Actual and predicted temperature data for North Soda Lake, California.

For temperature dataset, comparison of compression ratios obtained using proposed method and conventional method is shown in Figure 31. Using proposed method higher compression ratios are achieved. The maximum compression ratio using proposed method is 4.52 whereas using conventional method it is 2.95.

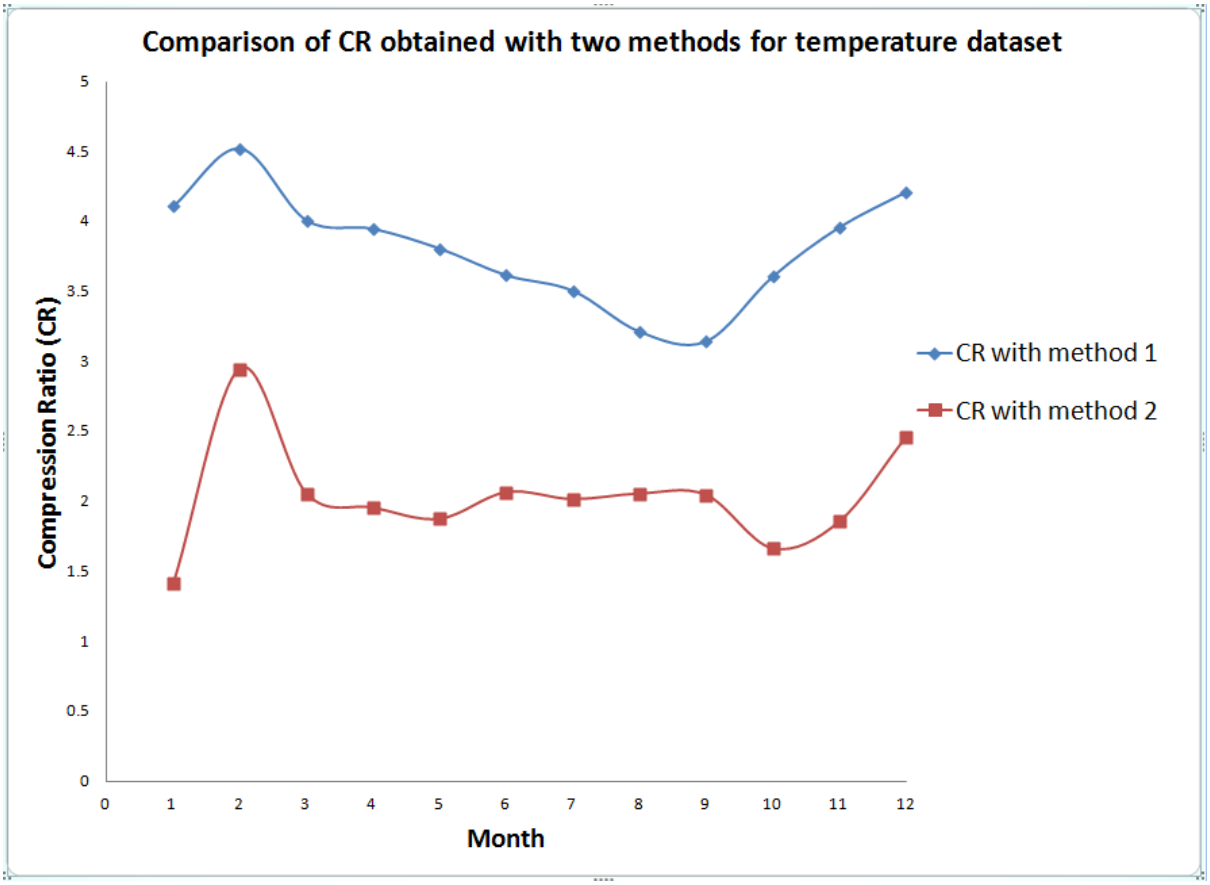

Figure 31. Comparison of compression ratio obtained with two methods for temperature data.

Table 17. Monthly compression ratios for the year 2012 for temperature data for North Soda Lake.

MONTH	UNCOMPRESSED BITS	COMPRESSED BITS (using method 1)	COMPRESSION RATIO (using method 1)	COMPRESSED BITS (using method 2)	COMPRESSION RATIO (using method 2)
JANUARY	10416	2534	4.11	7335	1.42
FEBRUARY	**9758**	**2159**	**4.52**	**3308**	**2.95**
MARCH	11160	2783	4.01	5417	2.06
APRIL	10800	2734	3.95	5510	1.96
MAY	11904	3124	3.81	6332	1.88
JUNE	11520	3182	3.62	5565	2.07
JULY	11904	3391	3.51	5893	2.02
AUGUST	11904	3697	3.22	5779	2.06
SEPTEMBER	11520	3657	3.15	5619	2.05
OCTOBER	11160	3091	3.61	6683	1.67
NOVEMBER	10800	2727	3.96	5806	1.86
DECEMBER	11160	2651	4.21	4537	2.46

CHAPTER 6

CONCLUSION AND RECOMMENDATIONS

This study presents an application of Artificial Neural Networks in predicting humidity and temperature data, thereby using this prediction for compression. In this thesis, ANN models were developed using multilayer feed forward neural network for prediction of data by employing past data inputs to predict the output. This chapter summarizes the goals and important results of this research. The conclusion identifies new research ideas that are listed as recommendations. Limitations and applications of this research are also described.

The first research question addressed was the variation of compression ratio using predictive analysis method or without it. One experiment was performed using predictive analysis method and another using conventional compression method. For predictive analysis method, the ANN model was used to predict the outputs which were then compared to the observed values to assess performance of the model and eventually the compression ratio was calculated. For the second method, conventional method of differential encoding followed by Huffman coding was adopted to calculate the compression ratio.

The second research question addressed was the variation of results with location of site. Experiments were performed on different locations to check variation in the results caused by the location of the site. The dependence of humidity and temperature on other parameters like pressure, solar radiation, month, etc was used as the basis of this study. Effect of location on the model was assessed by comparing two cases (a) minute interval dataset and (b) 10-minute interval dataset.

6.1. Conclusions

Compression of climate data is useful to scientists and researchers who deal with large climate databases. Artificial Neural Networks have proved to be a useful tool in prediction of data based on past values. In the past, most of the research for scientific data

compression has focused primarily on combining compression with data synthesis in order to increase throughput and conserve storage. This research uses different geographical parameters of a location as inputs to Artificial Neural Networks for prediction of humidity and temperature data. Key results of the research are listed below:

1. In case of humidity data, for a 1-minute dataset, maximum compression ratio of 6.14 and 2.66 is achieved using proposed and conventional method respectively; and for 10-minute dataset, maximum compression ratio of 5.26 and 2.24 is achieved using proposed and conventional method respectively.
2. For temperature dataset, maximum compression ratio obtained in the month of February, was 4.52 and 2.95 using predictive analysis method and conventional method respectively.
3. Data fluctuation is lower every 1 minute and the values are closer to each other one another which leads to better prediction and higher compression ratios when compared to data values every 10-minute or every hour. Such high compression ratio is useful in increasing storage capacity.

Thus, the proposed method is useful for lossless compression of climate data.

6.2. Recommendations

Based on the above conclusions, new questions and research ideas are identified. These could be investigated to extend the research work presented in this thesis. Suggestions to address those research ideas and extend this work include the following:

1. This research work could be extended by using other models of ANNs such as Cascade feed forward, Elman, Radial basis neural network models for prediction of data.
2. Proposed method could be applied to data average over a period of seven days.

3. Humidity and temperature model developed in this study could be implemented at other available locations with different sources of data as inputs to the model.
4. Proposed method can be applied to any dataset which consists of numerical data such as medical data, real estate data, natural gas price data, financial data, and also applicable in the field of geochemical modeling, and medical administration.

6.3. Limitations

Some of the limitations of the research are listed below:

1. To increase the performance of the prediction model, only the inputs related to output should be used.
2. Prediction model requires training of large dataset in order to predict the data accurately. In case of small dataset, prediction might not be accurate.
3. For estimation purposes, model requires the use of location specific data.

REFERENCES

[1] Blelloch, G. E. (2001). *Introduction to data compression.* Computer Science Department, Carnegie Mellon University.

[2] Sayood, K. (2012). *Introduction to data compression.* Newnes.

[3] Wolfram, S. (2002). *A new kind of science*. Champaign: Wolfram media, Volume 5. page 1069.

[4] Climate Data and data products. World Meteorological Organization. Retrieved from http://www.wmo.int/pages/themes/climate/climate_data_and_products.php

[5] Saupe, D., Hartenstein, H., & Wergen, W. (1996). *Compression of weather forecast data.* 12th International Conference on Interactive Information and Processing Systems (IIPS). American Meteorological Society, Atlanta (volume and issue number is unavailable)

[6] Steffen, C. E., & Wang, N. *Weather data compression*. NOAA Research-Forecast Systems Laboratory. Boulder, Colorado. Volume 4. no. 9. (The year is unavailable)

[7] Karim, S. A., Karim, B. A., Ismail, M. T., Hasan, M. K., & Sulaiman, J. (2010). *Compression of temperature data by using Daubechies wavelets*. In Proceedings of International Conference on Mathematical Sciences (ICMS2), Volume 30, pp. 726-734.

[8] Engelson, V., Fritzson, D., & Fritzson, P. (2000). *Lossless Compression of High-volume Numerical Data from Simulations.* In Data Compression Conference. page 574.

[9] Xie, X., & Qin, Q. (2009). *Fast lossless compression of seismic floating-point data.* In Information Technology and Applications (IFITA). International Forum on IEEE. Volume. 1, pp. 235-238.

[10] Yadav, A. K., & Chandel, S. S. (2014). *Solar radiation prediction using Artificial*

Neural Network techniques: A review. Renewable and Sustainable Energy Reviews, Volume 33, pp 772-781.

[11] Al-Alawi, S. M., & Al-Hinai, H. A. (1998). *An ANN-based approach for predicting global radiation in locations with no direct measurement instrumentation.* Renewable Energy. Volume 14. no. 1, pp 199-204.

[12] Sözen, A., Arcaklioğlu, E., Özalp, M., & Kanit, E. G. (2004). *Use of artificial neural networks for mapping of solar potential in Turkey*. Applied Energy. Volume 77. no. 3, pp 273-286.

[13] Sözen, A., Arcaklioğlu, E., & Ozalp, M. (2004). *Estimation of solar potential in Turkey by artificial neural networks using meteorological and geographical data.* Energy Conversion and Management. Volume 45. no. 18, pp 3033-3052.

[14] AbdAlKader, S. A., & AL-Allaf, O. N. (2011). *Back propagation neural network algorithm for forecasting soil temperatures considering many aspects: a comparison of different approaches.* In Proceedings of the 5th International Conference on Information Technology, Amman. pp. 11-13.

[15] Mummadisetty, B. C., Puri, A., Sharifahmadian, E., & Latifi, S. (2015). *Lossless Compression of Climate Data.* In Progress in Systems Engineering. Springer International Publishing. pp. 391-400.

[16] The sensor system, Nevada Climate Change Portal. (2015). Available from http://sensor.nevada.edu/NCCP/The%20Project/SENSOR%20System.aspx

[17] Dascalu, S., Harris Jr, F. C., McMahon Jr, M., Fritzinger, E., Strachan, S., & Kelley, R. (2014, June). An Overview of the Nevada Climate Change Portal. Proceedings of The 7th International Congress on Environmental Modelling and Software. Volume 1, pp. 75-82.

[18] Nevada Climate Change Portal. (2015). Available from

http://sensor.nevada.edu/NCCP/Default.aspx

[19] The Nevada Climate-ecohydrological Assessment Network (NevCAN), NCCP. (2015). Available from http://sensor.nevada.edu/NCCP/Climate%20Monitoring/Network.aspx

[20] McMahon Jr, M. J., Dascalu, S. M., Harris Jr, F. C., Strachan, S., & Biondi, F. (2011, January). Architecting climate change data infrastructure for Nevada. Advanced Information Systems Engineering Workshops, pp. 354-365. Springer Berlin Heidelberg.

[21] Equipment and Sensors: Detail, NevCAN, NCCP. (2015). Retrieved from http://sensor.nevada.edu/NCCP/Climate%20Monitoring/Equipment%20Details.aspx

[22] List of Sensors, NevCAN, NCCP. (2015). Retrieved from http://sensor.nevada.edu/NCCP/Climate%20Monitoring/Sensors.aspx

[23] Data Search Interface, Sensor Data, NCCP. (2015). Available from http://sensor.nevada.edu/Data%20Search/Silverlight%20Data%20Client.aspx

[24] Science for changing world, USGS. (2014). Available from http://www.usgs.gov/aboutusgs/default.asp

[25] Effects of Climatic Variability and Land Use on American Drylands, USGS, (2013). Available from http://esp.cr.usgs.gov/projects/sw/overview.html#problem

[26] Southwest Climate Impact Meteorological Stations (CLIM-MET), USGS. (2013) Available from http://esp.cr.usgs.gov/projects/sw/clim-met/intro.html

[27] CLIMET Station, USGS. (2013). Retrieved from http://esp.cr.usgs.gov/projects/sw/clim-met/anatomy/index.html

[28] Steinruecken, C. (2014). Lossless Data Compression (Doctoral dissertation, PhD thesis, Cavendish Laboratory).

[29] Bhaskar, A.K. (2001). Data compression techniques (MS thesis).

[30] Lelewer, D. A. & Hirschberg, D. S. (1987). Data compression. ACM Computing Surveys (CSUR), *19*(3), 261-296.

[31] Sayood, K. (2012). Introduction to data compression. Newnes.

[32] An introduction to digital graphics. (2011). [Web log post]. Retrieved from http://burnleyandpendlenews.blogspot.com/2011/01/introduction-to-digital-graphics.html (blog)

[33] History of Lossless data compression algorithms. (2014). Retrieved from http://ethw.org/History_of_Lossless_Data_Compression_Algorithms

[34] Wu, K., Otoo, E. J., & Shoshani, A. (2006). Optimizing bitmap indices with efficient compression. ACM Transactions on Database Systems (TODS), 31(1), 1-38.

[35] Huffman, D. A. (1952). A method for the construction of minimum redundancy codes. Proceedings of the IRE, 40(9), 1098-1101.

[36] Kodituwakku, S. R., & Amarasinghe, U. S. (2010). Comparison of lossless data compression algorithms for text data. Indian journal of computer science and engineering, 1(4), 416-425.

[37] Hajek, Milan (2005). *Neural networks*, University of KwaZulu-Natal.

[38] Zou, J., Han, Y., & So, S. S. (2009). *Overview of artificial neural networks*. In Artificial Neural Networks (pp. 14-22). Humana Press.

[39] Shebany, M. et al. (2014). *Artificial neural network: a brief overview*. In International. Journal of Engineering Research and Applications, Volume 4 (Issue 2), Version 1, (pp 07-12).

[40] Kozyrev, S. V. (2012). *Classification by ensembles of neural networks*. In P-Adic Numbers, Ultrametric Analysis, and Applications, (pp 27-33).

[41] Filippo, A. et al. (2013). *Artificial neural networks in medical diagnosis*. In Journal of applied biomedicine, (pp 47 - 58).

[42] Chrislb. (2010). *Diagram of a multi-layer feed forward artificial neural network.* Available from: https://commons.wikimedia.org/wiki/File:MultiLayerNeuralNetworkBigger_english.png#file

[43] Da Silva, I. N., Cagnon, J. Â., & Saggioro, N. J. (2013). *Recurrent neural network based approach for solving groundwater hydrology problems*. INTECH Open Access Publisher.

[44] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1988). *Learning representations by back-propagating errors.* Cognitive modelling, 5, Volume 3, (pp 533-536).

[45] Mu Sigma. (2014). *Neural network*. Retrieved from http://www.mu-sigma.com/analytics/thought_leadership/cafe-cerebral-neural-network.html

CURRICULUM VITAE

Astha Puri

5188 Wilbur Street, Las Vegas, Nevada- 89119
Tel: (702) 945-9093
Email: astha.puri029@gmail.com

Education

University of Nevada Las Vegas, Nevada 2015
Master of Science in Electrical and Computer Engineering GPA: 3.93

Punjab Technical University, Jalandhar, India 2011
Bachelor of Science in Electrical Engineering 78.7%

Professional Experience

University of Nevada, Las Vegas **Jan 2014 - Dec 2015**
Research Assistant: Solar-Water-Environment Nexus Project, Nevada
- Worked as a Research Assistant for Data Analysis of Nevada Climate change portal.
- Responsible for climate data compression algorithms.

Wipro Technologies, India **Oct 2011 - Dec 2013**
Project Engineer: Nokia Siemens Networks Project, Helsinki, Finland, US
- Designed algorithms for Sales Order, Purchase Order, Invoices, Goods Receipt and Credit notes.
- Responsible for developing and maintaining Modular Logistics System (MLS) application based on Oracle Forms Application covering the critical areas of Enterprise Resource Planning.
- Responsibilities involved daily monitoring of batches and programs.
- Month end and Year end activities involving closure of user interfaces and Monthly report creations.

Intern- Employee Information System, Wipro Technologies, India.
- Designed, developed and maintained an Employee Information System.
- Responsibilities involved checking duplicate entries and generating reports.
- Developed and maintained the project.

Skills

- Data Processing & Modeling- MATLAB
- Languages- C, C++, SQL, PL\SQL, No SQL, Mongo DB, JAVA
- Web Technologies- HTML, Java Script, PHP

Publication

- Astha Puri, Bharath Chandra Mummadisetty, Ershad Sharifahmadian, Shahram Latifi "A Hybrid Method Lossless Compression of Solar Radiation Data Using Neural Networks", International Journal of Communications,Network and System Sciences,2015.
- Astha Puri, Ershad Sharifahmadian and Shahram Latifi, "A comparison of hyperspectral image compression methods", International Journal of Computer and Electrical Engineering, (IJCEE), Volume 6, Issue 6, Pages 493-500, December 2014.
- Astha Puri, Bharath Chandra Mummadisetty, Ershad Sharifahmadian,Shahram Latifi, "Lossless Compression of Climate data",Proceedings of Twenty-Third International Conference on Systems Engineering,pp 391-400,Las Vegas,Aug 2014

Professional Certifications

- Fundamental Engineer (FE) exam, Nevada Board of Engineers & Land Surveyors 2014