Lossless Compression of Humidity and Precipitation Data

Bharath Chandra Mummadisetty, Astha Puri, and Shahram Latifi

***Abstract*—Given the explosive growth of data that needs to be transmitted and stored, there is a necessity to focus on developing better transmission and storage technologies. The main goal of this paper is to develop the best compression methods for climate data. By using data compression, significant reduction in the bits to encode the climate data can be accomplished without loss of any important information. In this paper, humidity and precipitation data are considered for compression. The methodology is based on differential encoding wherein the prediction of the sample being encoded is obtained according to the output of an ANN (multilayer perceptron) whose inputs are time, month, temperature, pressure, incoming shortwave radiation, incoming long wave radiation, outgoing shortwave radiation, outgoing long wave radiation and solar radiation data for humidity and time, month, temperature, humidity, incoming short wave radiation, outgoing shortwave radiation, incoming long wave radiation, outgoing long wave radiation, and solar radiation data for precipitation. The ANN model uses 3 layers and 27 neurons for prediction of humidity data and 2 layers and 20 neurons for precipitation data. The highest compression ratio for precipitation data is 8.96 which is observed for the month of October and for humidity data the highest compression ratio is 6.92 and it is observed in the month of June.**

***Index Terms*—Compression, compression ratio, artificial neural networks, climate data.**

I. Introduction

We have been witnessing a revolution in the past few years in the way the communication works and it is still in the evolution stage. This includes the fast growing internet; the growth of mobile and video communication. Data compression is one of the means to make this transformation possible. It would be impractical to provide better clarity for communication through cell phones; it would be difficult to have images, audios, videos on websites without data compression. With the help of data compression one can listen to music on music player, watch movies of DVD, make long distance calls etc. The data to be compressed can be text in a file, numbers, audio signals, and images generated due to different processes. More information generated and used is in digital form and it is represented by bytes of data. And the number of bytes required to represent multimedia data can be huge [1].

Data compression is also known as source coding which means encoding information with fewer bits than the original representation. This suitable encoding can be possible by recognizing the patter and structure that under lies in the data. Data can be anything from letter, symbols, words from text file to images and videos, it can also be data generated from other processes. There two ways in which data compression can be classified i.e lossy and lossless compression. In lossless compression, the original information can be retrieved back without any loss of information whereas in lossy compression, some loss is incurred when the reconstruction take place. So original information is not retrieved in lossy compression. Depending on the application one can decide to go for lossy or lossless compression technique. For example, when we consider text compression one cannot chose to go for lossy compression as we know that in lossy compression some loss is incurred. As text involves characters and if any character is missed or there is a discrepancy then it makes a huge difference with respect to the original data. In the applications where exact representation for original data is not required, one can go for lossy compression. For example, during audio compression lossy compression is favorable as accurate value of each sample is not required. The loss of information in this can be tolerated to different amounts depending on the quality of the reconstructed signal [2].

The values of different climate parameters, such as surface temperature, precipitation and humidity are calculated at each grid point over each time step to predict their future values. Therefore, there is a demand to develop alternative ways of predicting these data [3]. Currently, finer resolution of the grid and shorter time intervals are adopted in climate simulation. With this more accurate simulation of the climate data is obtained. Thus, the size of the climate modeling output is expanding exponentially along with the increase in the grid resolution and time steps. The huge volume of climate modeling data is bringing tough challenges to the storage system and network system for data archiving and data sharing respectively. Data compression is an effective way to mitigate these challenges by compressing the data size, reducing the amount of data to be stored and transferred. However, climate data should be archived losslessly to maintain numerical stability for the climate models [4].

The most common way to measure the efficiency of a compression technique is by computing the compression ratio. It is defined as the ratio of number of bits required to represent the data before compression to the number of bits required to represent the data after compression [1]. Climate data is very important for research studies at the same time it is voluminous. Climate data is ever growing, given the explosive growth of data that needs to be transmitted and stored; there is a necessity to focus on developing better techniques of the same.

ANNs are models used to gauge or approximate functions

Manuscript received July 15, 2015; revised November 20, 2015. This work was supported in part by the National Science Foundation (NSF) grant #EPS-IIA-1301726 and in part by the Nevada NASA EPSCOR Space Grant.

The authors are with the Department of Electrical and Computer Engineering, UNLV, Las Vegas, USA (e-mail: bharath.mummadisetty@gmail.com, astha.puri029@gmail.com, Shahram.Latifi@unlv.edu).

which have large number of inputs and outputs. They are inspired by biological neural networks of animals. ANNs are interconnected system of neurons which communicate with each other. The neurons are connected to each other with the help of links which bear weights and they can be tuned based on experience and making them capable to learn. ANNs can solve wide variety of problems as it can learn from data like applications in rule-based programming, speech recognition, computer vision etc. ANNs are also used to model populations and environment which change frequently. They can be built using hardware and also through software. There are different types of ANNs such as feedforward neural network, radial basis function network, elman neural network, cascade feed forward neural network, physical neural network etc. We have considered multilayer perceptron(MLP) for our research. MLP model consists of multiple layers of nodes, with each layer connected to the next one usually in a feed forward way. It consists of either three or more than three layers: input layer, one or more hidden layers, and an output layer of nonlinearly-activating nodes. Each node in one layer connects with a certain weight to every node in the following layer. This model maps sets of input data onto a set of appropriate outputs. Each neuron in one layer has direct connections to the neurons of the subsequent layer [5].

The ANN models have been used in various researches such as in climate data. The ANN models used different geographical parameters of a location as inputs for the prediction of solar radiation as discussed in [6]. In the study undertaken by AbdAlKader and AL-Allaf, back propagation neural network models were developed to predict the day soil temperature for the present day by using various previous day meteorological variables in Nineveh-Iraq. After ANN training the models gave a close prediction to the actual values [7]. In the research work done by Tamer Khatib, Azah Mohamed, K. Sopian, M. Mahmoud , prediction of hourly solar radiation values for Kuala Lumpur was performed. This prediction was performed using the GRNN, FFNN, CFNN, and ELMNN artificial neural networks. Prediction results show that GRNN has a higher efficacy compared to the other proposed networks. The FFNN and CFNN are still efficient at predicting solar radiation but do not predict well in poor radiation conditions such as the first and final hour of the solar day. The ELMNN was the worst at predicting the solar radiation among the proposed methods [3]. In the research work done by Bharath Chandra Mummadsietty and Astha Puri, the problem of lossless, offline compression of climate data was addressed. They proposed a method for compression solar radiation, photo synthetically active radiation, and data logger power system voltage data using combination of differential encoding and Huffman coding [2].

The source for data collection is Nevada climate change portal, also known as NCCP. This portal provides data download facility for climatic, hydrological, ecological, and hardware data. Different types of sensors are places are placed in Snake and Sheep mountain range. There are totally 12 sites namely sheep range mojave desert shrub, sheep range subalpine, sheep range black brush, sheep range pinyon-juniper, sheep range montane, snake range east salt desert shrub, snake range west pinyon-juniper, snake range east sagebrush, snake range west subalpine, snake range west sagebrush, snake range west montane, and snake range east subalpine. Fig. 1 shows the location of these sites. Each site has data values for 25 parameters in a time gap of 1- minute and 10-minute. For this study, we deal with precipitation data and humidity data for the years 2013 and 2014 [8].

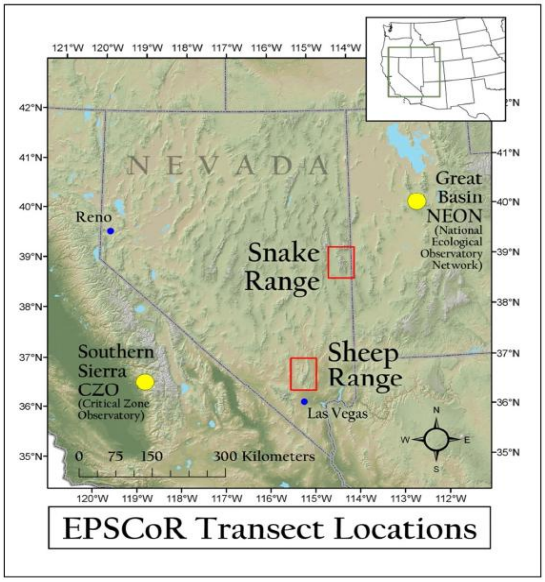

Fig. 1. Location of the sites [4].

The rest of the paper is organized as follows. Section II describes the proposed method. Results and discussion is presented in Section III, whereas Section IV concludes the paper.

II. Proposed Method

A. Method for Precipitation Data

The precipitation, temperature, humidity, incoming radiation data, outgoing radiation data, and solar radiation data are downloaded from Nevada climate change portal for the year 2013. All the inputs and outputs are minute wise time series data. There are total of 1440 data points per day. The training data fed to the ANN is of the year 2013.The inputs are stacked together as 2 dimensional matrix of 8 × 525600 having a total of 4204800 data points. Here precipitation is measured in mm(millimeter) and temperature in deg(degrees). The incoming radiation, outgoing radiation, and solar radiation are measured in w/m^2. There are 8 inputs for this method namely time, month, temperature, humidity, incoming short wave radiation, outgoing shortwave radiation, incoming long wave radiation, outgoing long wave radiation, and solar radiation data. The input data is for the year 2013 and the output a 1 dimensional matrix having 44640 data points. The inputs and the output are presented to the ANN. In this case, we have used multilayer perceptron neural network for training the given data set. After multiple iterations, the best performance i.e the mean square error of 10.1 is obtained. To obtain this performance, we have used ANN with 2 layers and 20 neurons in each layer. In the next step, we download the input data for year 2014 from Nevada climate change portal. This data is presented to the model trained above and then predicted values for precipitation data for the year 2014 are obtained. Then the precipitation data for the

year 2014 is downloaded from the data portal. Since the precipitation data contains floating point numbers, all the data points are multiplied by 10 to make both inputs and outputs as integers. In this way it will be beneficial for pre-processing the data. Now the difference between actual and predicted data is computed. When the performance of MSE is good then the difference will be small which is great for the compression. On the data obtained from the previous stage, differential encoding is applied which will help in increasing the probability of occurrence of a symbol. Upon this data, Huffman coding is applied which will encode the data with fewer bits thus giving lossless data compression of precipitation data.

As part of decompression, the first step is to decode the Huffman code. Now the first number obtained from decoding process is kept as the first data point for the next stage. We add this number to the second number from the decoded data which will be the second data point for the next stage and this process will continue till the last data point of the decoded data. Next, each data point from this stage is added to the predicted data. After this each of the data point is divided by 10 to give us the actual data. Thus the total process of lossless compression and decompression is completed for precipitation data. The proposed method is shown in Fig. 2.

B. Method for Humidity Data

For humidity data, we have considered 9 inputs namely time, month, temperature, pressure, incoming shortwave radiation, incoming long wave radiation, outgoing shortwave radiation, outgoing long wave radiation and solar radiation data. The dataset for the year 2013 is fed to the ANN as part of training. Output is the humidity data for the year 2014. All of this data is downloaded from Nevada climate change portal. Here the input is a 2 Dimensional matrix of 9 ×525600 having 4730400 data points. In this case we have used ANN with 3 layers and 27 neurons present in each layer. Here humidity is measured in deg-m, pressure in % and rest of the parameters has the same units as mentioned in the procedure for compression of precipitation data. The output data is 1 a dimensional matrix of 44640 data points.

C. Flowchart

After numerous iterations, MSE of 44 is obtained. This is achieved with combination of 3 layers and 27 neurons in each layer. Humidity data is also floating point data with 2 digits after the decimal point. The procedure used for humidity is similar as the method used for compression of precipitation data , the only difference is after data is predicted by ANN both the actual and predicted data have to be multiplied by 100 before the difference is computed. During decompression the data obtained before the final stage has to be divided by 100 to retrieve back the original data losslessly.

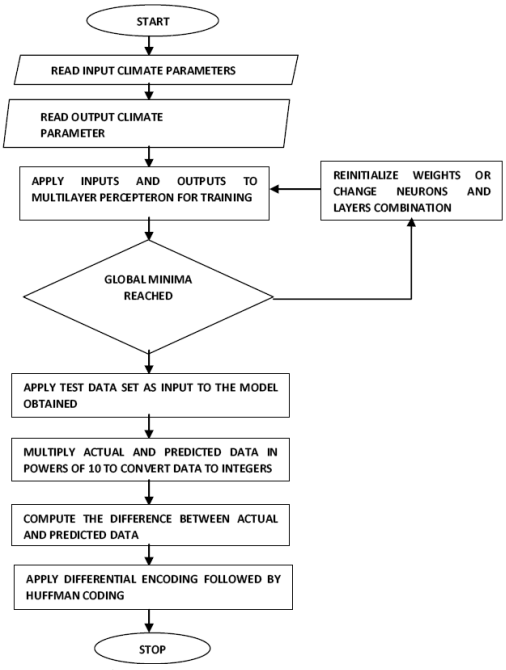

Fig. 2. A flowchart describing the proposed method.

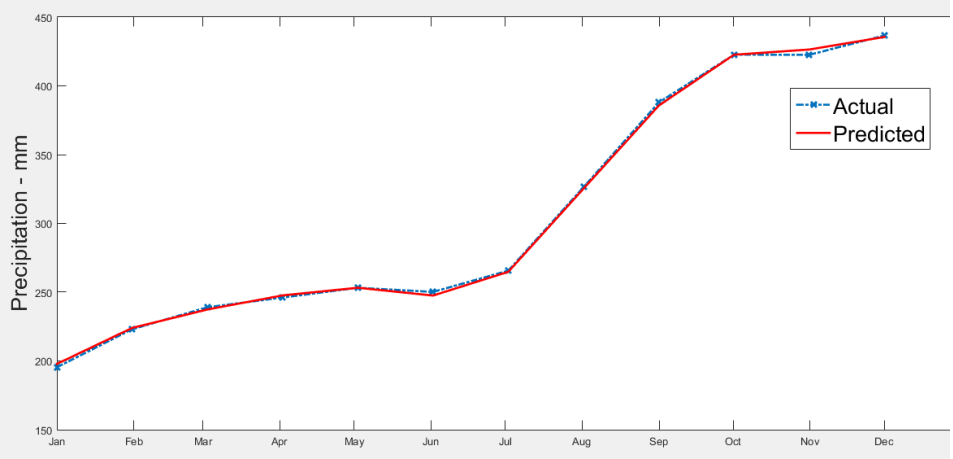

Fig. 3. Actual and predicted data for Precipitation for year 2014.

III. Results and Discussion

The results obtained for the procedures are given in this section. The compression ratios are given in Table I and Table II. The graphs for actual and predicted data are given in the Fig. 3 and Fig. 4.

Fig. 3 and Fig. 4 show the plots of actual and predicted precipitation and humidity data respectively. From the plots, it can be observed that the actual and predicted data are very close to each other which is good for compression of data. This is possible because ANNs are great at prediction of climate data.

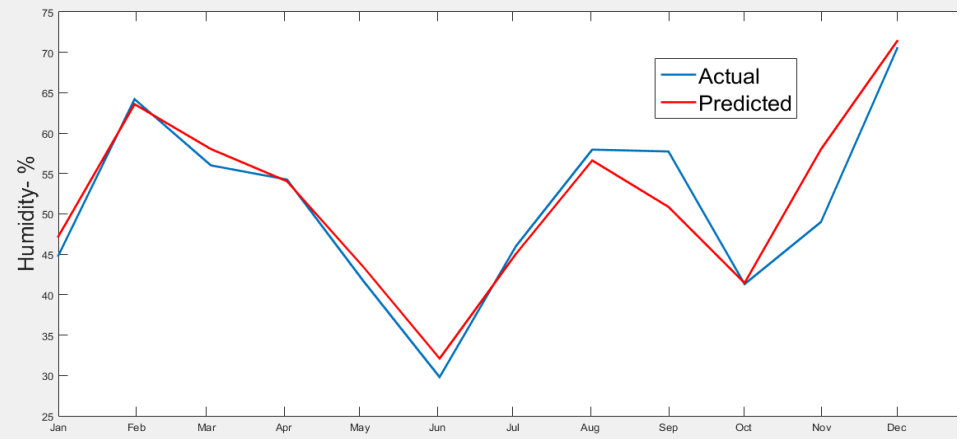

Fig. 4. Actual and predicted data for humidity for year 2014.

TABLE I: COMPRESSION RATIO FOR PRECIPITATION DATA FOR THE YEAR 2014

MONTH	UNCOMPRESSED BITS	COMPRESSED BITS	COMPRESSION RATIO
JANUARY	758880	85608	8.86
FEBRUARY	685440	79467	8.62
MARCH	758880	91529	8.29
APRIL	734400	87308	8.41
MAY	758880	90276	8.40
JUNE	758880	90090	8.42
JULY	803520	95594	8.40
AUGUST	803520	99824	8.04
SEPTEMBER	777600	103394	7.52
OCTOBER	**803520**	**89579**	**8.96**
NOVEMBER	777600	97382	7.98
DECEMBER	803520	102828	7.81

TABLE II: COMPRESSION RATIO FOR HUMIDITY DATA FOR THE YEAR 2014

MONTH	UNCOMPRESSED BITS	COMPRESSED BITS	COMPRESSION RATIO
JANUARY	758880	140794	5.39
FEBRUARY	685440	111092	6.17
MARCH	758880	142915	5.31
APRIL	734400	147766	4.97
MAY	758880	145379	5.22
JUNE	**758880**	**109664**	**6.92**
JULY	803520	156326	5.14
AUGUST	803520	147977	5.43
SEPTEMBER	777600	143733	5.41
OCTOBER	803520	153051	5.25
NOVEMBER	777600	139605	5.57
DECEMBER	803520	138061	5.82

IV. CONCLUSION

In this paper, we proposed lossless compression methods to compress the humidity and precipitation data extracted from Nevada climate change data portal. The methodology adopted the uses multilayer perceptron neural network for prediction of climate data. For performance metrics, compression ratio and root mean square error were calculated. The highest compression ratio for precipitation data was 8.96 which was observed for the month of October and for humidity data the highest compression ratio was 6.92 observed in the month of June. Thus, our study shows that with the proposed method climate data variables can be compressed with a high compression ratio which is useful for increasing storage capacity.

REFERENCES

[1] K. Sayood, *Introduction to Data Compression*, Second Edition.
[2] B. C. Mummadisetty, A. Puri, and S. Latifi, "Lossless compression of climate data," *Progress in Systems Engineering, Springer* International Publishing, pp. 391-400, 2015.
[3] T. Khatib, A. Mohamed, K. Sopian, and M. Mahmoud, "Assessment of artificial neural networks for hourly solar radiation prediction," *International Journal of Photoenergy*, 2012.
[4] L. Songbin, X. Huang, Y. Ni, H. Fu, and G. Yang, "A high performance compression method for climate data," in *Proc. IEEE International Symposium on Parallel and Distributed Processing with Applications*, 2014, pp. 68-77.
[5] B. C. Mummadisetty, A. Puri, and S. Latifi, "A hybrid method for compression of solar radiation data using neural networks," *International Journal of Communications, Network and System Sciences*, vol. 8, no. 6, p. 217, 2015.
[6] A. K. Yadav and S. S. Chandel, "Solar radiation prediction using artificial neural network techniques: A Review," *Renewable and Sustainable Energy Reviews*, vol. 33, pp. 772-781.
[7] S. A. Abd AlKader and N. A. Al-Allaf, "Backpropagation neural network algorithm for forecasting soil temperatures considering many aspects: A comparison of different approaches," in *Proc. the 5th International Conference on Information Technology*, Amman, 2011.
[8] S. Dascalu, F. Harris, and E. Fritzinger, "An overview of the nevada climate change portal," in *Proc. the 7th International Congress on Environmental Modelling and Software*, vol. 1, 2014.

Bharath Chandra Mummadisetty received his bachelor degree in telecommunication engineering from MVJ College of Engineering, Bangalore, India in 2010. He is currently pursuing his master in electrical and computer engineering in University of Nevada, Las Vegas (UNLV). He has been working in NSF EPSCoR project as a research assistant for the past one and half years. His research interests include machine learning, artificial neural networks, data compression and software engineering.

Astha Puri received her BS degree in electrical engineering from Chitkara University, Punjab, India in 2011. Currently, she is pursuing her master degree in the Department of Electrical and Computer Engineering, University of Nevada, Las Vegas (UNLV), Las Vegas, Nevada. Her research work focuses on climate data compression and hyperspectral images.

Shahram Latifi received his M.S. and Ph.D. degrees in electrical and computer engineering from Louisiana State University in 1986 and 1989, respectively. He is currently a professor of electrical and computer engineering and the co-director of Center for Information Technology and Algorithms at the University of Nevada, Las Vegas. Dr. Latifi has taught courses and performed research in various areas including image processing and document analysis, data compression, remote sensing, biometrics, security and computer networks. He has authored/co-authored over 200 technical articles in various journals and conferences. Dr. Latifi was an IEEE distinguished speaker (1997-2000), an associate editor of the IEEE Transactions on Computers (1999-2006), co-founder and general chair of the IEEE Int'l Conf. on Information Technology (ITCC 2000-2004) and founder and General Chair of Int'l Conf. On Information Technology-New Generations ITNG (2005-2015). He has also served on the editorial board of several international journals.