J 9.7   **INVESTIGATION OF DATA COMPRESSION TECHNIQUES APPLIED TO AWIPS DATASETS**

Ning Wang*, Sean Madine*, and Renate Brummer*
NOAA Research–Forecast Systems Laboratory
Boulder, Colorado

*[In collaboration with the Cooperative Institute for Research in the Atmosphere (CIRA)
Colorado State University, Fort Collins, Colorado]

1. INTRODUCTION

The current AWIPS system has an incoming data volume of 5-8 GB per day. This volume will continue to increase with the development of higher resolution numerical models and more frequently available observations. With the maturation of network computing technology and the need for distributed meteorological workstations, this large volume of data must be frequently transmitted over the Internet or other communication channels (satellite links, for example) to the workstations located remotely. So it becomes increasingly important to apply appropriate data compression techniques to reduce the volume of data, and it also makes it possible to deliver and store the large volume of data on the remote meteorological workstations.

There are many commercially available packages for data compression. Unfortunately, they are mostly designed for textual content (e.g., gzip, pkzip, compress), natural pictures (e.g., JPEG, GIF), or motion pictures (mpeg2 encoder etc.). They are tuned for these specific data and for speed, and may not achieve desired compression performance or meet some important requirements for meteorological datasets. Also, the data compression research, especially in the field of lossy compression has made huge progress in the last decade. New compression techniques, mostly transform-based, have been developed providing more options and challenges for applying the compression techniques to scientific data sets.

In this article, we investigate the use of the compression techniques for various types of meteorological datasets. The natures of different compression techniques and the data to be compressed are first discussed. Then we present the results from our experiments with real data and conclude with some remarks about existing problems and future direction.

Corresponding author address:
Ning Wang, NOAA/FSL, Boulder, CO
phone: 303-497-6704; email: wang@fsl.noaa.gov

2. DATA COMPRESSION SCHEMES

The objective of this investigation is to identify the appropriate compression methods for various meteorological datasets, specifically AWIPS datasets, and to determine the feasibility of using data compression in distributed network meteorological workstations.

Data compression in general can be classified into two big categories: lossless compression and lossy compression. Lossless compression achieves data size reduction without information loss. Lossy compression reduces data size with controlled fidelity loss. For this fidelity loss, we can do one of the following: specify the compression ratio (the ratio between the size of original data and that of the compressed data) and let the encoder minimize the error under some predefined error metric (usually L^2), or give an error limit and let the encoder minimize the compressed data size. In general, lossy data compression achieves a higher compression ratio.

Ideally, we want to compress all data in the lossless fashion. However, there is a limit on how much data size reduction lossless compression can achieve. In practice, to achieve the highest compression possible, we need to decide what kind of loss each data type can accept and how much compression is needed.

For the lossless data compression, the highest compression is usually achieved by arithmetic coding method with finite-context models (Bell et al 1990). Obviously, lossless compression does not allow us to specify the compression ratio. For an encoder, all one can do is to tune certain parameters to fit specific data. There is a limit on the compression performance.

The research in lossy data compression has been quite active since the introduction of the wavelet transform, an orthogonal transform with its basis function localized in both "time" and "frequency". Along with an appropriate quantization scheme, it achieves superior compression compared with other transform-based compression schemes. Even with the same PSNR value (Peak Signal and Noise Ratio, a measure of the fidelity of the reconstructed data, defined in terms of mean square error), wavelet compressed pictures show a superior visual result.

The largest meteorological datasets involve the output of numerical forecast models. The size of a typical parameter of a mesoscale model output is around 80MB. Lossless compression can achieve a compression ratio between 1.1 : 1 and 1.5 : 1. Many parameters have fields that are nicely smooth and a small amount of error is usually acceptable. We are using wavelet transform-based compression for these datasets. First, datasets are specially preprocessed so that we can take advan-tage of the correlation of the data in time and space. Then a separable multidimensional wavelet transform is performed to concentrate the energy and reduce the correlation. It is followed by a standard quantization procedure and an entropy encoding procedure with an order-2 arithmetic encoder. With minimum error, we can achieve compression ratios from 40:1 to 300:1.

The next category is the weather imagery data. Among them, there are three major types: satellite pictures, radar images, and model grids rendered in color image forms. For radar images, where every pixel is rather important, there is no room for any errors. On the other hand, because of the nature of data, radar images usually consist of "sparse" echoes with relatively low depth (color) resolution, and therefore, are quite compressible with lossless compression encoder. We compress all radar images using the standard lossless image encoder (e.g. GIF encoder). In recent experiments, we applied an arithmetic encoder to the radar images and achieved higher compression. Since the lossless compression method does not work well for satellite pictures, we use the wavelet compression for them. The criteria we use for the acceptable loss are rather subjective: the reconstruction error must be visually unnoticeable and the reconstructed pictures must be "meteorologically useful."

The last category is the vector graphic files. These are the graphical products of observations, or model output. They can only be compressed with lossless compression. Among these graphical products, most of them are small in size and therefore need not be compressed. Only a few of them are large; for example, surface observation plot products have data sizes around few hundred kilobytes per frame. For those datasets we apply a high order entropy encoder, which results in compression ratios of 5:1 or better.

3. EXPERIMENTAL RESULTS

The FX-Net project has been an ideal test bed for compression of different AWIPS datasets. We apply wavelet transform-based compression to all satellite image products that are transmitted for FX-Net clients. With very little visual loss, we achieve 7:1, 15:1 and 50:1 compression ratios for visible, infrared, and water vapor channel images. More details are provided in Madine and Wang (1999).

Radar images are compressed in the lossless way. The following table compares the compression performance

of the standard GIF encoder with arithmetic encoder.

	Clear Sky	Storm Mode
GIF Encoder	34.9 : 1	8.7 : 1
Arithmetic Encoder	54.4 : 1	12.6 : 1

The volume grid data we tested are the output of the MM5 mesoscale model. The two parameters selected are temperature and relative humidity at different levels above the ground. Temperature represents a relatively smooth field, and has a lot of space correlation between adjacent grid points, it is relatively easy to compress. In contrast, relative humidity involves many rapid changes, or high frequency components. From the quantization point of view, its transformed coefficients have much higher geometric mean, therefore they are much harder to compress (Jayant and Noll 1984). Figures 1a and 1b present the average absolute errors at different levels of the temperature field (Fig. 1a) and the relative humidity field (Fig. 1b) with different compression ratios applied.
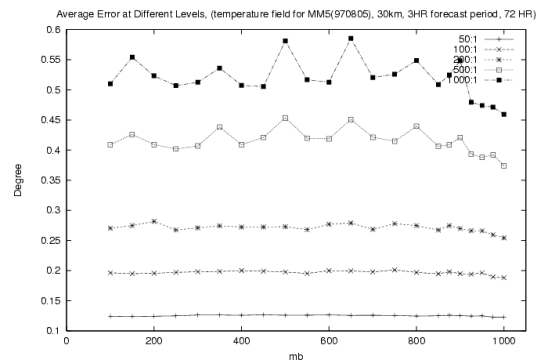

Figure 1a: Average (absolute) error at different vertical levels for a temperature field with different compression ratios.
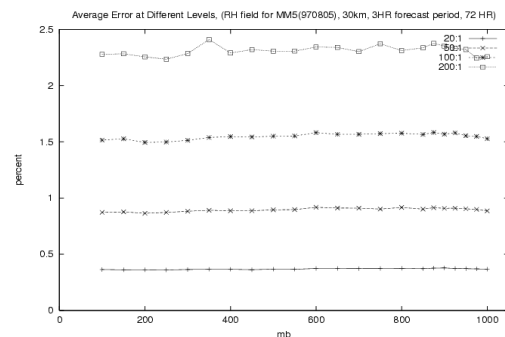

Figure 1 b: Average (absolute) error at different vertical levels for a relative humidity field with different compression ratios.
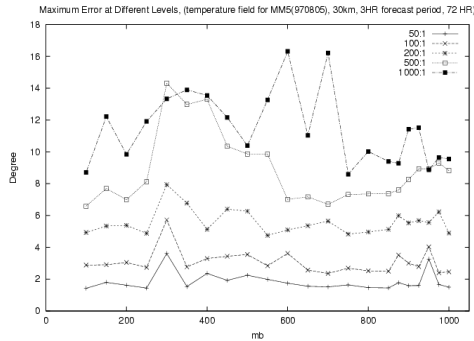
Figure 2 a: Maximum error for the temperature field considering different compression ratios.
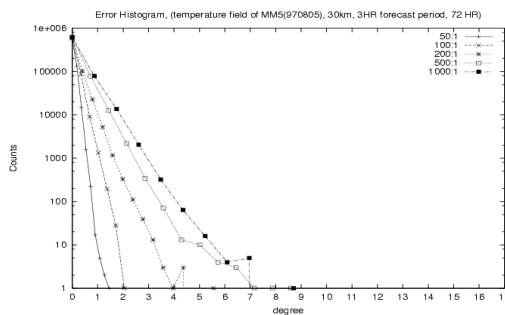

Figure 2 b: Error histogram (logarithmic scale) for the temperature field considering different compression ratios.

We make two observations here. First, with compression ratios chosen quite high (between 50:1 and 200:1 for temperature and between 20:1 and 50:1 for relative humidity), the average errors turn out to be very small (Fig. 1a and 1b). The magnitudes of these errors vary only a little between levels. In contrast, the maximum errors vary in a much larger range as can be seen in Figure 2a. They seem to peak at the levels close to the surface and tropopause. Second, from the error histogram we learn that large magnitude maximum errors (Fig. 2a) have only few counts in the error histogram (Fig. 2b). The distribution seems to approach a Poisson distribution with a small mean value. The practical conclusion of the experimental tests for the temperature field is that the original 80MB file size (MM5, 30-km resolution, 3 hourly forecasts, 72 hour forecast period) can be compressed down to 400KB by applying a compression factor of 200:1, with very little loss in fidelity.

4. CONCLUDING REMARKS

We have discussed and presented compression techniques for meteorological datasets. The results from the experiments indicate that data compression technology is very useful and promising in its application to the meteorological datasets. The use of these compression techniques in the FX-Net Project (Wang and Madine 1998) is very successful. The compression scheme developed for grid-type datasets has been evaluated and tested. High compression ratios have been achieved with minimum fidelity loss for those smooth datasets. However, there are still many issues concerning the practical use of data compression. We need further analyses about the impact of the loss of information. There are many demands on the lossless data compression with higher compression ratios. Further investigation may be needed in the area of transform-based lossless data compression. Another important aspect is the computation time involved in the compression and decompression procedures. For a compression scheme to work for datasets with ever-increasing size, we have to consider the computation complexity. For transform-based compression algorithms, more effort is needed to efficiently perform the quantization and entropy coding.

5. REFERENCES

Bell, T.C., J.G. Cleary, and I.H. Witten, 1990: *Text Compression*, Prentice Hall, Englewood Cliffs, NJ.

Jayant, N.S., P. Noll, 1984: *Digital Coding of Waveforms: Principles and Applications to Speech and Video*. Prentice Hall Signal Processing Series, A.V. Oppenheim, ed., Englewood Cliffs, NJ.

Madine, S., and N. Wang, 1999: Delivery of meteorological products to an Internet client workstation. *15th Inter. Conf. on Interactive Information and Processing Systems for Meteorology, Oceanography, and Hydrology*, Dallas, TX, Amer. Meteor. Soc., 356–359.

Wang, N., and S. Madine, 1998: FX-Net: A Java-based Internet client interface to the WFO-Advanced workstation. *14th Int. Conf. on Interactive Information and Processing Systems for Meteorology, Oceanography, and Hydrology*, Phoenix, AZ, Amer. Meteor. Soc., 427–429.