



A Comparison Study of Data Compression Techniques for the Battle-Scale Forecast Model (BFM) Grids



Ning Wang and Cliff Matsumoto
DoD Center for Geosciences/Atmospheric Research, CIRA/Colorado State University, Fort Collins, Colorado

Objective

This study focuses on the comparison of three data compression techniques applied to the Battle-scale Forecast Model (BFM) grids—two developed by CIRA/CSU and one implemented jointly by the University of Texas at El Paso (UTEP) and the Army Research Lab (ARL) at White Sands, NM. The Root Mean Square Error (RMSE) and Maximum Absolute Error (MAE) achieved by the three different techniques were compared at several fixed bit rates. In addition, the expected computational complexities for these different compression schemes in the context of an operational setting are also discussed.

Dod Relevance

- Increased data volume (model grids as well as satellite imagery/sensor data) requires efficient compression method for **data transmission and storage / archival**.
- Varied data applications drive the need for **easily specifying and controlling the desired precision (error tolerance)** of the compressed data.
- Varied users require simplified and **"portable" user interface** and application platform.

Battle-scale Forecast Model (BFM) Dataset

The data set for the BFM consists of three wind components (U, V, and W), potential temperature (THET), pressure (PR), and water vapor mixing ratio (WV). The primary grid structure for each variable is a 129x129x64x2 4-D data set, representing the x, y, z 3-D spatial grids at two forecast periods. In the work performed by [1], all original floating point BFM data is first converted to fixed-point (16 bits) data prior to any steps in the data compression procedure. For this data compression evaluation, CIRA-developed new codecs maintain the floating point data format throughout the encoding and decoding procedures.

Strategies for Data Compression

• ARL/UTEP codec

A 1-D Karhunen-Loeve (KL) transform is carried out as the pre-process step for all fields along the z direction before the data set is fed to the JPEG2000 encoder, which then encodes each horizontal field as a 2-D image. The KL transform de-correlates the data in the vertical dimension so that the JPEG2000 encoder can better perform its encoding process. To optimize the compression, ARL/UTEP codec allocates bit rate for each slice of the grid based on its variances.

• CIRA/CSU codec

A three-dimensional wavelet transform is applied to a folded 4-D grid of each variable. We believe this is a more computationally efficient way to transform the data set, and is very comparable in terms of the ability to de-correlate the original dataset's correlation in all direction. The 3D wavelet transform in general is defined as,

$$F(s, \vec{x}) = \int_{R^3} f(\vec{x}') \psi_{(s, \vec{x})}(\vec{x}') d\vec{x}', \vec{x}, \vec{x}' \in R^3.$$

For the CIRA/CSU data compression technique, a discrete form of the transform is used, and the transform coefficients at dyadic points are obtained through applying filter banks along z and x,y directions, separately.

In lossy data compression, the fidelity of the reconstructed datasets are often measured in l^p norm in error vector space which is defined as:

$$\|E\|_p = \left(\sum_{i=1}^n (f_{org}[i] - f_{rec}[i])^p \right)^{1/p},$$

with l^∞ being defined as:

$$\|E\|_\infty = \max_{1 \leq i \leq n} |f_{org}[i] - f_{rec}[i]|,$$

where n is the number of discrete data points. $\|E\|_2$ is often used since it measures the Euclidean distance between two data sets, hence the average, or overall fidelity of the reconstructed data set. For scientific data, it is also desirable to know what precision the reconstructed data set maintains, i.e. $\|E\|_q$. In many cases, either one or both of them are minimized. While $\|E\|_2$ is invariant under orthogonal transform (or approximately invariant under biorthogonal transform), $\|E\|_q$ is not. This presents a challenge for most transform-based data compression methods, where quantization procedures are usually done in the transformed domain.

An quantization algorithm is developed and implemented to achieve high fidelity data compression with respect to both error metrics. It is a two-step quantization procedure. The first step of the algorithm aims at reducing the error measured in l^2 norm, and then the second step further reduces the l^∞ norm error. Two different quantization schemes are used in these steps. The optimal / suboptimal bits allocation for the two steps is achieved by using the statistical information obtained from the training data sets. This scheme is the improved algorithm described in [2].

Evaluation and Comparison Results

The three codecs evaluated in this comparison study are the codecs with a 3D wavelet transform and single quantizer ('W3d w/sq'), a 3D wavelet transform with double quantizers ('W3d w/dq'), and the JPEG 2000 plus KLT preprocessing codec developed by UTEP and ARL [1] ('ARL'). Tables 1 and 2 show the Root Mean Square Error (RMSE) and the Maximum Absolute Error (MAE) for all model variables at 0.9 bit rate (~35:1 compression), for all three codecs. Units are deg K for THET, m/sec for U, V, and W, millibars for PR, and g/kg for WV.

Table 1. RMSE for all codecs at 0.9 bit rate (~35:1 compression)

	THET	U	V	W	PR	WV
ARL	0.03118	0.00514	0.01332	0.00033	0.00463	0.00413
W3d w/dq	0.01402	0.00562	0.00749	0.00041	0.00264	0.00281
W3d w/sq	0.01595	0.00624	0.00835	0.00045	0.00369	0.00315

Table 2. MAE for all codecs at 0.9 bit rate (~35:1 compression)

	THET	U	V	W	PR	WV
ARL	0.94772	0.11308	0.17760	0.00845	0.03861	0.06201
W3d w/dq	0.06097	0.02400	0.00310	0.00279	0.01401	0.01360
W3d w/sq	0.11581	0.07134	0.10456	0.00518	0.05768	0.02928

Figures 1 and 2 show the graphical comparisons of the average errors and maximum absolute errors for the potential temperature field at different bit rates.

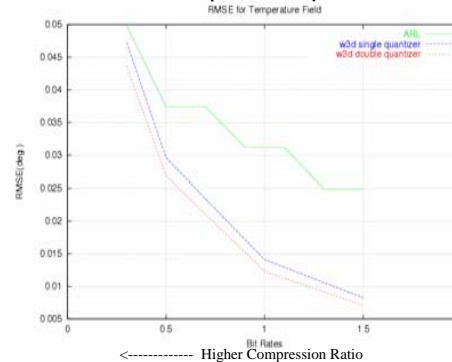


Fig. 1. RMSE (deg.) for the potential temperature field for the three codecs.

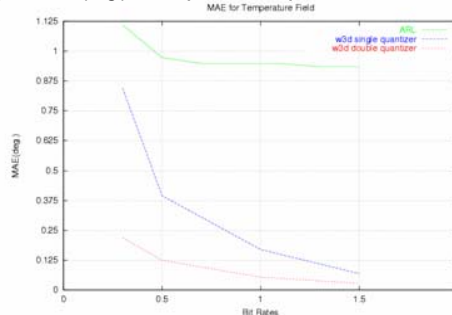


Fig. 2. MAE (deg.) for the potential temperature field for the three codecs.

Tables 3 and 4 list the computation time for the encoding and decoding procedures for the CIRA/CSU codec with double quantizations at the various bit rates for all model variables. Since only the Java version of the ARL codec (which runs substantially slower) was available for this study, a fair comparison with its use of CPU time was not possible for this study. The machine used for this test was a Dell Precision 420 with 900 MHz Intel Pentium III (coppermine) CPU.

Table 3. CPU time for encoder (in seconds)

	THET	U	V	W	PR	WV
0.3 Bit rate	5.70	5.66	5.57	5.62	5.65	5.61
0.5 Bit rate	5.72	5.82	5.78	5.79	5.65	5.82
1.0 Bit rate	6.14	6.19	6.08	6.15	6.12	6.09
1.5 Bit rate	6.51	6.46	5.91	6.63	6.54	6.28

Table 4. CPU time for decoder (in seconds)

	THET	U	V	W	PR	WV
0.3 Bit rate	3.38	3.36	3.33	3.35	3.34	3.39
0.5 Bit rate	3.50	3.54	3.47	3.48	3.45	2.86
1.0 Bit rate	3.78	3.83	3.75	3.86	3.81	3.86
1.5 Bit rate	4.03	4.05	3.98	4.21	4.08	4.18

Conclusions

• The two CIRA/CSU codecs demonstrate better overall performance in both l^2 error metric and l^∞ error metric for all variables rather consistently, with only few exceptions, when compared to the ARL codec (see [3] for the entire comparison study results for all parameters). This indicates that the CIRA/CSU quantization scheme is suitable for the data set that was tested. This should be the case for other high-resolution numerical model output fields, as well.

• The CIRA/CSU codec with double quantization appears quite effective in reducing the maximum absolute error. Therefore, it should be considered as a viable alternative coding format to the current coding format (e.g., GRIB I and II) for high-resolution gridded data. The computation cost of the codec with double quantizer is a little higher than the one with single quantizer, but still quite acceptable.

• The entire model data set can be compressed in less than 40 seconds (Table 3) and decompressed in less than 25 seconds (Table 4) on a 900Mhz Pentium III machine.

Operational Implications

• Grid data compression (quasi-lossless) with precision control (user-defined acceptable maximum/average error) of 20:1 to 100:1 achievable.

• Average storage and transmission time reduced 95% to 98%.

• Earlier effort demonstrated usable visible satellite imagery compression of 20:1 to 50:1.

• Technique is already a core component for operational platform supporting fire weather, air quality, university weather labs, and other operational agencies.

• An intelligent 'push' system for entire process—from data ingest, through data compression and data delivery to remote clients—called Compression Relay Management System now being assembled.

Acknowledgement

We gratefully acknowledge the funding support for this research by the DoD Center for Geosciences/Atmospheric Research at Colorado State University under Army Research Laboratory Cooperative Agreement DAAD19-02-2-0005.

References

- [1] Koshelova, O., A. Aguirre, S. D. Cabrera and E. Vidal, Jr., *Assessment of KLT and bit-allocation of JPEG 2000 to the Battlescale Forecast Meteorological Data*, 2003. IEEE Geoscience and Remote Sensing Symposium, Toulouse, France.
- [2] Wang, N. and R. Brummer, *Experiment of a wavelet-based compression technique with precision control*, 2003. Proc. of 93rd American Meteorological Soc. 19th Conf. on IIPS for Meteorology, Oceanography, and Hydrology, Long Beach, CA, February 2003.
- [3] Wang, N. and C. Matsumoto, *A comparison study of data compression techniques for the Battle-scale Forecast Model grids - Phase II*, 2005. Progress Report for the Army Research Lab, White Sands, NM, May 2005.