A Comparative Study of Lossless Compression Algorithms on MODIS data

Srikanth Gottipati[a] and Jamal Goddard[a] and Michael Grossberg[a] and Irina Gladkova[a]

[a]CCNY, NOAA/CREST, 138th Street and Convent Avenue, New York, NY 10031, USA

ABSTRACT

This paper reports a comparative study of lossless compression algorithms for MODIS data. MODIS, The Moderate Resolution Imaging Spectroradiometer, is a 36 band Visible and IR multispectral imager aboard the Terra and Aqua satellites, having spatial resolution ranging from 0.250 to 1 kilometer and spectral resolution ranging from 0.405 -0.420 to 4.482-4.549 microns. MODIS data rates are 10.6 Mbps (peak daytime); and 6.1 Mbps (orbital average). Faced with such an enormous volume of data on a current generation imager, this study provides a comparison of current compression algorithms as a baseline for future work. The Hierarchical Data Format (HDF) is standard format selected for data archiving and distribution within the Earth Observing System Data and Information System (EOSDIS). Currently this system handles over one terabyte of data daily, and this volume continues to increase over time. With growing satellite Earth science multispectral imager volume data compression, it becomes increasingly important to evaluate which compression algorithms are most appropriate for data management in transmission and archiving. This comparative compression study uses a wide range standard implementations of the leading lossless compression algorithms. Examples include image compression algorithms such as PNG and JPEG2000, and widely-used file compression formats such as BZIP2 and 7z. This study includes a comparison with the Consultative Committee for Space Data Systems (CCSDS) most recent recommended compression standard. by a significant margin.

Acknowledgments

Research sponsored by NOAA/NESDIS under Roger Heymann (OSD), Tim Schmit (STAR) Compression Group.

Keywords: compression, MODIS

1. INTRODUCTION

High resolution multi-spectral imagers are becoming increasingly important tools for understanding and monitoring the earth. Future NOAA missions such as GOES-R will include improved imagers which will provide a rich stream of scientific data. The rich stream of data comming from next generation imager must be transmitted wirelessly back to earth over channels with severely limited bandwidth. Even after data are received they must be archived, and distributed world wide. This makes lossless compression of the data essential.

As a proxy for next generation imagers we have focused on the Moderate Resolution Imaging Spectroradiometer (MODIS). MODIS is a 36 band Visible and IR multispectral imager which is currently deployed on both the Terra and Aqua satellites. The uncompressed imaging data consists of digital measurement counts, calibration information and meta-data such as telemetry of the satellite. In this study we focus on compression of the imaging data since the size of the other data is negligible in comparison. Both processed and unprocessed data are distributed. Since other forms may be derived from it, we will focus on the compression of the unprocessed Level 1A digital counts.

There are number of lossless compression algorithms available. We considered a number of widely used algorithms which are relatively fast, achieve high compression results. We look at a representative collection of general compression algorithms such as Bzip2 and 7zip, and image compression algorithms such ad PNG and Jpeg2000. We also considered important to look at algorithms for which there are open implementations

Further author information: (Send correspondence to)
Michael Grossberg: E-mail:grossberg@cs.ccny.cuny.edu, Telephone: 212-650-6166

Table 1. MODIS data set attributes

Resolution	No of sensors	No of rows	No of columns	No of channels	Memory (bits)	Memory (%)
250m	40	2080*4	1400*4	2	1490944000	39%
500m	20	2080*2	1400*2	5	931840000	24%
1km day	10	2080	1400	14	652288000	17%
1km night	10	2080	1400	17	792064000	20%

available. This availability insures that the compression algorithms we consider here, may be used globally for transmission, distribution and archiving of imager data.

One concern in any evaluation is the variability of the data. Evaluation by the examination of a small number of samples, or exclusively focusing on the mean of the data is typically insufficient for use in planing engineering requirements. In this work we have used the approach of stratified sampling to address this. We have broken the data into non-data dependent sample bins and sampled from each. This improves accuracy by reducing bias that may inadvertently be present in the data that has been made available. It also allows us to examine the dependence of the data on the factors we use to make the stratification. We conclude by showing that the estimates across the sampling are consisitant giving high confidence in the reliability of the estimates for future data.

2. MODIS DATA

The MODerate resolution Imaging Spectroradiometer (MODIS) is a key instrument aboard the Terra (EOS AM) and Aqua (EOS PM) polar satellites. Terra's orbit around the Earth is timed so that it passes from north to south across the equator in the morning, while Aqua passes south to north over the equator in the afternoon. The MODIS instrument provides high radiometric sensitivity (12 bit) in 36 spectral bands ranging in wavelength from 0.4 μm to 14.4 μm. These 36 distinct spectral bands are divided into four separate Focal Plane Assemblies (FPA): Visible (VIS), Near Infrared (NIR), Short- and Mid-Wave Infrared (SWIR/MWIR), and Long-Wave Infrared (LWIR). Each FPA focuses light onto a certain section of detector pixels, which are relatively large, ranging from 135 μm to 540 μm square. The large number and variety of detector pixels are what make the wide variety of MODIS data possible. When light hits a detector pixel, it will generate a distinct signal depending on the type of light it is sensitive to. The signals that the pixels generate are what scientists process and study to learn about Earth's land surfaces, water surfaces, and atmosphere. There are 10 detector elements along track for each of the 1 km bands, 20 for each of the 500 m bands, and 40 for the 250 m bands. A more detailed overview of the MODIS data attributes for various band types are shown in Table 1.

The instruments record digital counts at 12 bit precision in each band. Due to the three different spatial resolutions for the different spectral bands, the data is not distributed evenly across spectral. The visible and near IR band at 250m resolution accounts for 39% of the total data size, as is illustrated in Figure 1. Perhaps not surprisingly the 5 bands which make up the next highest resolution at 500m make up the next largest portion of the data. From this we see that from the point of view of compression, accounting for spatial relationships within MODIS data would seem essential for good compression. Nevertheless we will see that despite this intuition algorithms such as Bzip2 and 7zip which do not use 2d spatial relationships, perform only slightly worse than wavelet based algorithms.

3. COMPRESSION ALGORITHMS

3.1 Tiff

TIFF stands for Tagged Image File Format is used for storing images and is well known for high color depth images. TIFF was developed by the Albus Corporation in 1986 and is a variable-resolution format. TIFF files

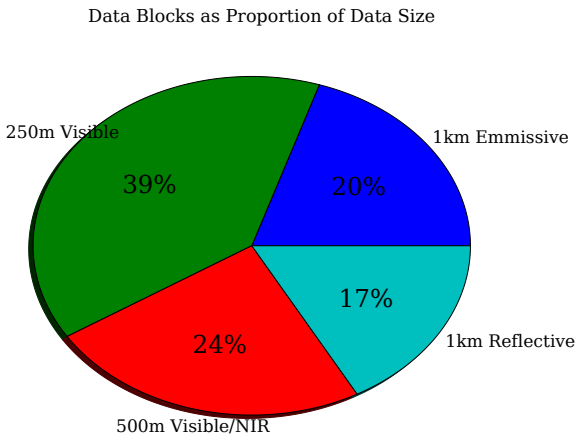

Figure 1. A pie graph showing the relative sizes of the data blocks within the MODIS Granule. The 2 bands of 250m Resolution Visible make up the largest portion of the data at 39%.

has the option to be compressed. TIFF files can be grey scale, RGB full color or several other classes. For the purpose of this experiment Lempel-Ziv-Welch usually referred to as LZW was the compression type used to compress the image files. To compress the LZW algorithm[1,2] builds a translation table.

3.2 Png

Portable Network Graphics (PNG) is a lossless compression format. Since it is fully lossless PNG can supports up to 48-bit truecolor or 16-bit grayscale file storage without losing any quality. PNG uses Adam7 algorithm. A PNG image is broken into seven layers, each layer is defined by placing an 8x8 pattern over the layer. The subimages are then stored in the PNG file in numerical order. Those layers are stored in the PNG file and Adam7 uses seven passes to reconstruct the image.

3.3 7zip

7-Zip is an open source file archiver designed by Igor Pavlov.[3] To compress file 7-Zip uses a combination of filters that can be preprocessors, compression algorithms, or encryption filters. The main compression is done using LZMA compression. LZMA algorithm was also developed by Igor Pavlov. LZMA uses an enhanced LZ77 algorithm[4] along with a form of arithmetic coding known as a range coder.

3.4 Bzip

Bzip2[5] is a lossless data compression algorithm that was developed by Julian Seward. Unlike rar and zip, Bzip2 is a data compressor, not an archiver. Bzip2 uses Burrows-Wheeler transform to change character sequences into strings of identical letters, then uses a move-to-front transform, and finishes using Huffman coding.

3.5 CCSDS-IDC-V2.0.6

The Consultative Committee on Space Data Systems (CCSDS) Image Data Compression (IDC) was developed for use on space instruments and uses an extended rice algorithm. CCSDS came up with the algorithm to help reduce data transmission time, reduce storage requirement, as well as to reduce the size of the data. This compression algorithm allows for real time processing with space electronics, it support 4 - 16 bit image compression, and can be adjusted to output lossy to lossless compression formats. CCSDS IDC[6] is a wavelet encoder that processes the coefficients from the residual subands with a special algorithm which makes it different than your normal wavelet encoders.

3.6 Jpeg2000

Jpeg2000 was created by the Joint Photographic Experts Group committee in the year 2000. Jpeg2000 uses multi-level discrete wavelet transform, that uses scalar quantization and block-based arithmetic coding to compress images. Jpeg2000 can achieve lossy to lossless compression. Its algorithm differs from standard Jpeg's which uses a discrete cosine transform to do its compression. We use the JasPer library[7] for implementing Jpeg2000.

4. STRATIFIED SAMPLING

For this project Stratified sampling was done. Stratified sampling is completed through several parts. To begin you take the initial population and divide it into none overlapping groups. Each member of the population is then separated and placed into one of the groups. This process continues until every member is in a group. No one member can be in two groups. Then a random or systematic sampling is done on each group. For this project random sampling was done and because of uneven sampling a weighted mean was calculated for each subsample.

4.1 Implementation

The MODIS data was downloaded from NASA. The data was divided into groups by season and hemisphere based in the time, data, and location of the satellite provided by NASA. There is a flag in the HDF indicating wheather the data observed the surface in daylight, at night or has points in both catagories called "Day", "Night", and "Both". Together this group structures describes all possibilities for acquired MODIS Level 1A data. To begin the script first chooses a random year from a list containing 3 years (2004 - 2006). After a year is chosen then a season is chosen. Once a season is selected then a day within that season is then selected. This is used to genterate a list of critia we need for this sample. The HDF files are then filtered for those criteria and within each of the files matching the criteria random files are selected using the rand function from the python V2.5 math library.

The compression of the file was also done using a python script which called linux command line programs. The two main computers running the script were a Pentium 4 3.60 MHZ computer, with 2GBs of ram and a linux OS system and a Dell computer with an Intel Xeon 3 MHZ processor, with 2GBs of ram and an Linux OS system. The linux distribution we used was Kubuntu "Feisty Fawn" and all the compression packages and libraries were those included or updated from that distribution. We used both the HDF4 and HDF5 libraries. Most of the HDF4 libraries were used extensively. The main function that we used was EDP which we allowed us to output header information. We also used it to dump the selected band (EV 250m, EV 500m, EV 1km day, EV 1km night) out as a binary file which is then converted to a pgm file using the linux command line function rawtopgm. Those compression types are TIFF,PGM,7zip,Bzip2,CCSDS IDC, and Jasper (Jpeg2000). The compression was all done using linux command line arguments. TIFF and PNG compressions were done using the netpbm library which contains pnmtotiff and pnmtopng. (With pnmtotiff the -lzw switch was used). 7z was used for 7zip compression, bzip2 for Bzip2, BWT and DWT was used for CCSDS IDC compression.

Before CCSDS IDC algorithm could be done each pgm file needed to be padded so that the length and the width were powers of two. Once compression is done then the size of the new compressed file is collected and the compression ratio calculated. All the data is then saved into a csv file and then feed into a database. Different data is collected from the header file including Day, Night, or Both (which is the time of day), also determine whether the granule is in the northern or southern hemisphere through the longitude and latitude.

5. EXPERIMENTAL EVALUATION

We show the data broken down by data block and hemisphere in Figures 2,3, 4,5, 6 and by data block and season in Figures 7,8, 9,10, 11. Finally we present our overall results in Figure 12 by data block.

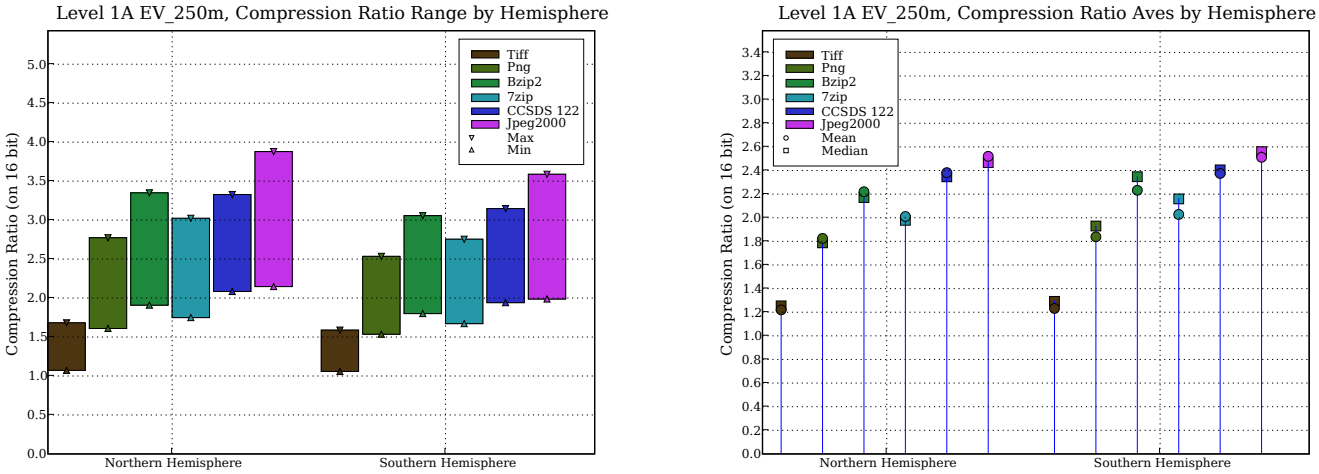

Figure 2. The graph on the left shows the range of compression ratios using six standard algorithms on the 2 bands of 250m resolution where the the northern and southern hemispheres were considered separately. The graph on the right shows means and medians (averages) for the algorithms. Neither the ranges nor the averages show much significant difference between the hemispheres. Jpeg2000 has the widest range sometime achieving as much as 3.9 to 1 compression and is the best performer.

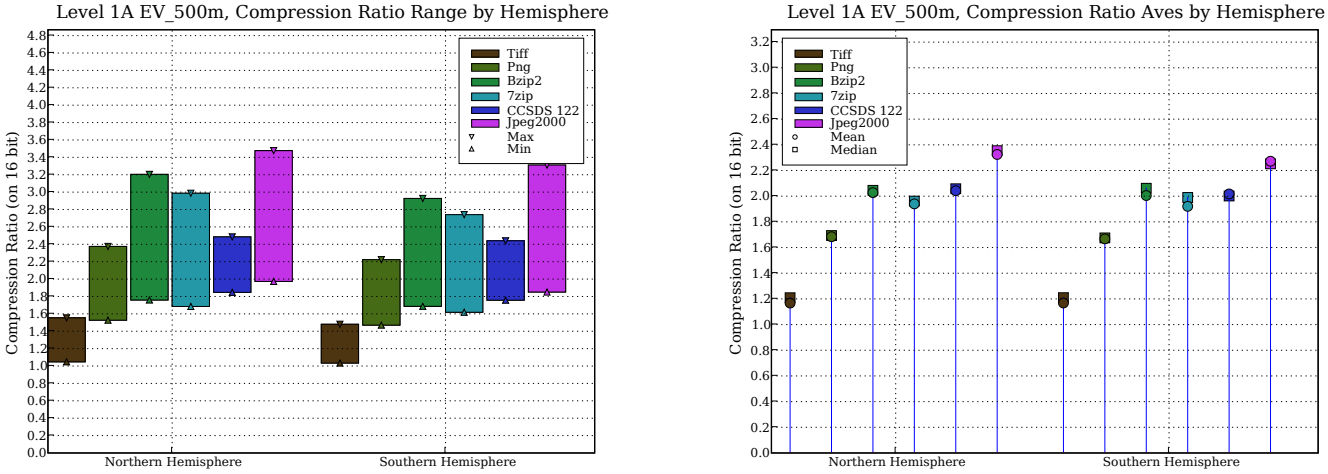

Figure 3. Ranges and averages for 5 compression algorithms applied to the 5 channel 500m resolution MODIS bands. Despite the differences in the characteristics of the 500m and the 250m resolution bands, the compression data is quite similar. Jpeg2000 compression is the again the overall winner with bzip2 and the new CCSDS standard comming quite close. As with the higher partial resolution channels we see little overall difference between northern and southern hemisphere data.

6. CONCLUSION

We have conducted an evaluation of 6 popular compression algorithms on MODIS 1A data. We have used a stratified sampling technique in order to get robust estimates. We have shown that our evaluations are consistent both across seasons and by hemisphere. We have reproduced our earlier results form entropy estimates that there is great variability in the 1km reflective data, and that both the 250m and 500m are the most difficult parts to compress. The fact that our results that there is little difference in compression ratios between data

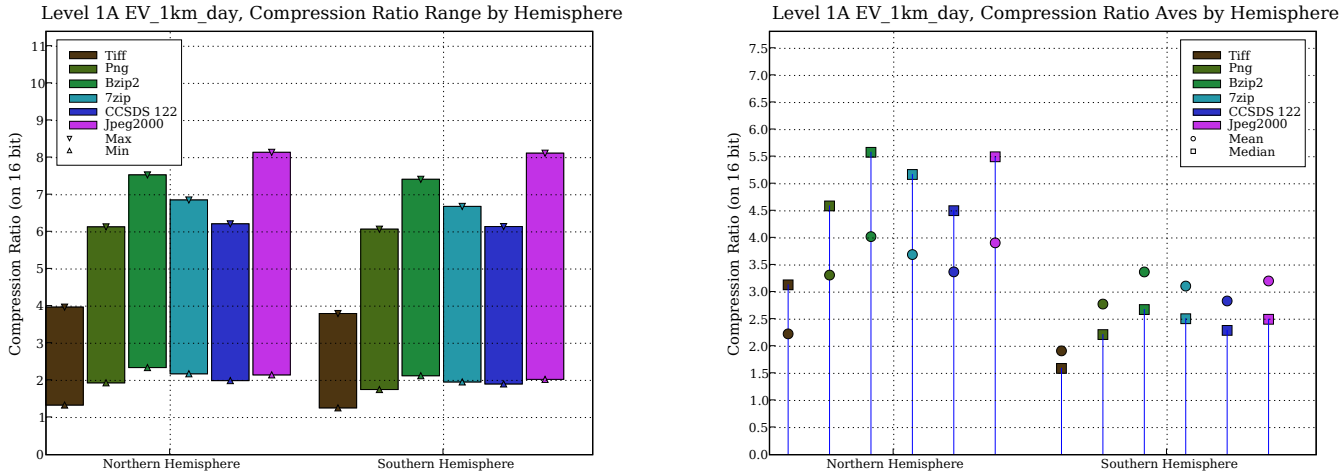

Figure 4. Graphs showing the ranges and average compression ratios for the reflective 1km resolution MODIS bands for northern and southern hemispheres. These bands were by far the most variable. Some of the channels would often saturate completely, which accounts for the extremely high maximum compression ratio. The difference between the median and means and the variation between the southern and northern hemispheres are all signs of the great variations within the data. Even so the relative performance of the algorithms is unchanged. Surprisingly Bzip2 is able to beat JPEG2000 on average. This is because when the data is saturated and nearly constant, Bzip2 is able to exploit that better than the wavelet based algorithms such as CCSDS and JPEG2000.

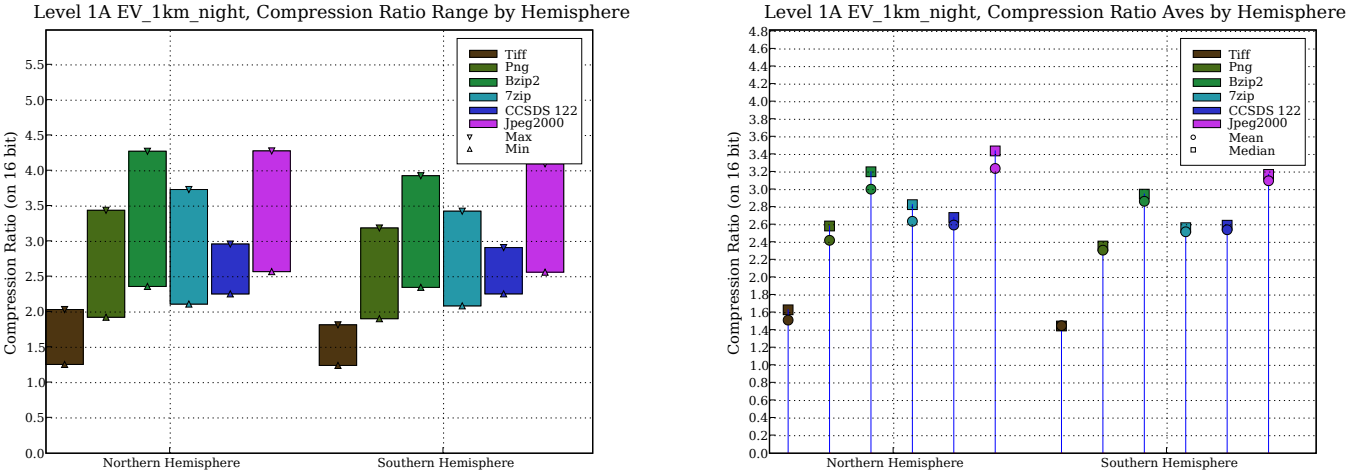

Figure 5. Graphs showing the ranges and average compression ratios for the em-missive 1km resolution MODIS bands for the northern and southern hemispheres. The statistics and properties of the 17 MODIS 1km (night) em-missive bands have statistics quite distinct from the reflective bands. Thus it perhaps comes as less of a surprise that the Jpeg2000 and CCSDS algorithms do not perform as well as they do on the other more image-like bands. Overall Jpeg2000 does somewhat better than other algorithms but, like on the 1km reflective data, BZip2 does remarkably well. This data provides the greatest challenge for imaging algorithms although it makes up a relatively small portion of the overall granule. As one might expect there is little difference between the northern and southern hemisphere data.

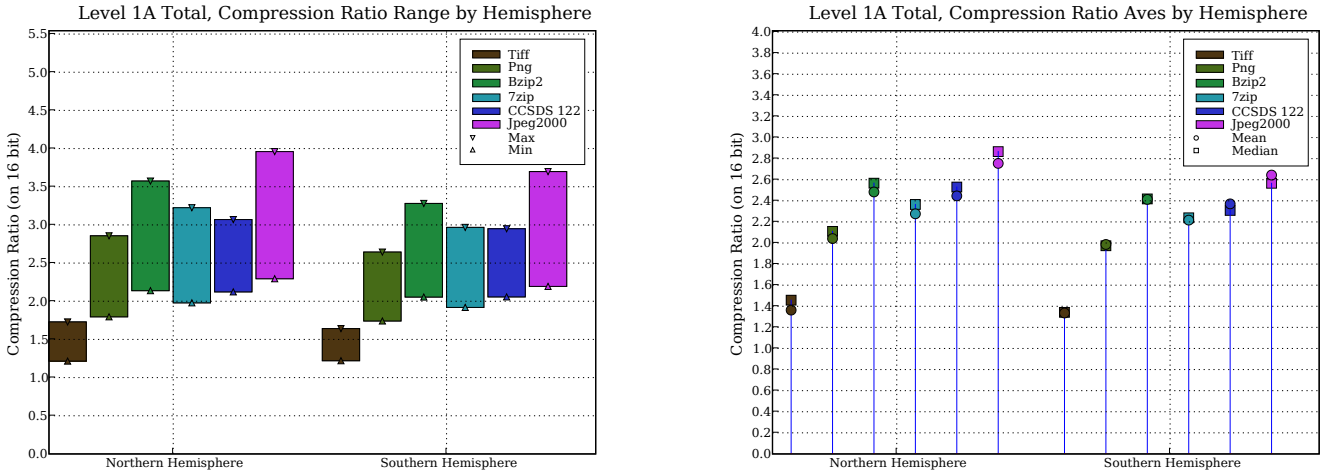

Figure 6. These graphs show the total ranges and average compression ratios for each of the algorithms both for the northern and southern hemispheres. The total results look qualitatively similar to the results for the 250m resolution. There is little difference between the northern and southern hemispheric data. Also while Bzip2 is behind Jpeg2000 it remains competitive.

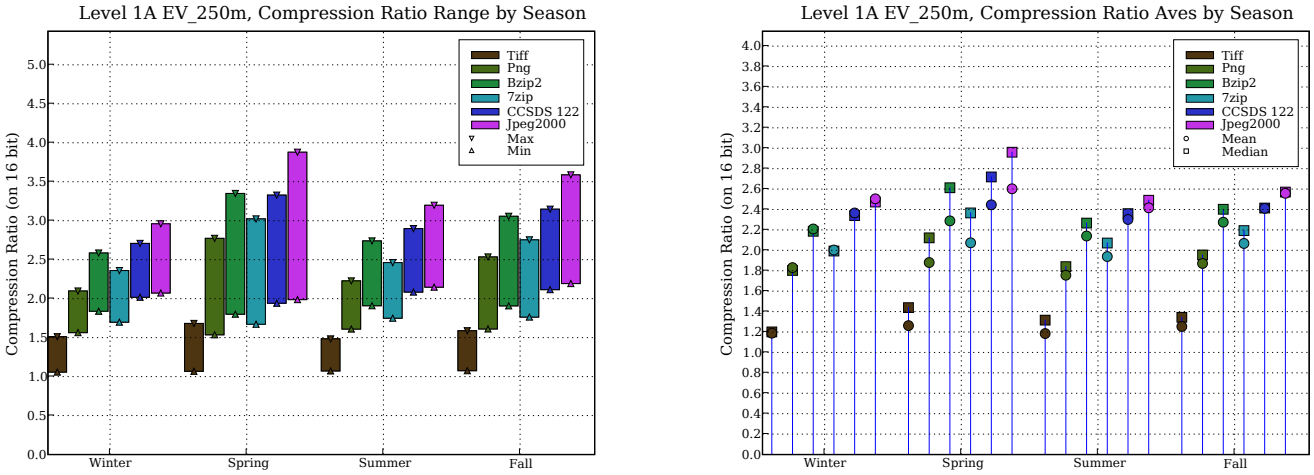

Figure 7. Graphs showing ranges and averages of compression ratios of 250m resolution data for each season (corrected for hemisphere). We see that there is little systematic seasonal variation of the compression visible in the 250m resolution data. However, there are indications of greater variability in Spring and Fall.

acquired in northern and southern hemispheres, may be interpreted as indicating that the atmospheric rather than the surface component has a much more significant impact on compression. The overall best performer was JPEG2000. This was expected given that we are not considering spectral correlations and that 250m and 500m resolution bands are similar to conventional image data. It was also expected that the CCSDS performs only slightly worse JPEG2000 since it is based on a similar wavelet filter bank. The excellent performance of Bzip2 was a surprise. It bests JPEG2000 in a few cases and does nearly as well overall. It does so despite the fact that it is not specifically an image compression algorithm. This seems to indicate that the compression performance could be improved on this data set by combining elements from both algorithms.

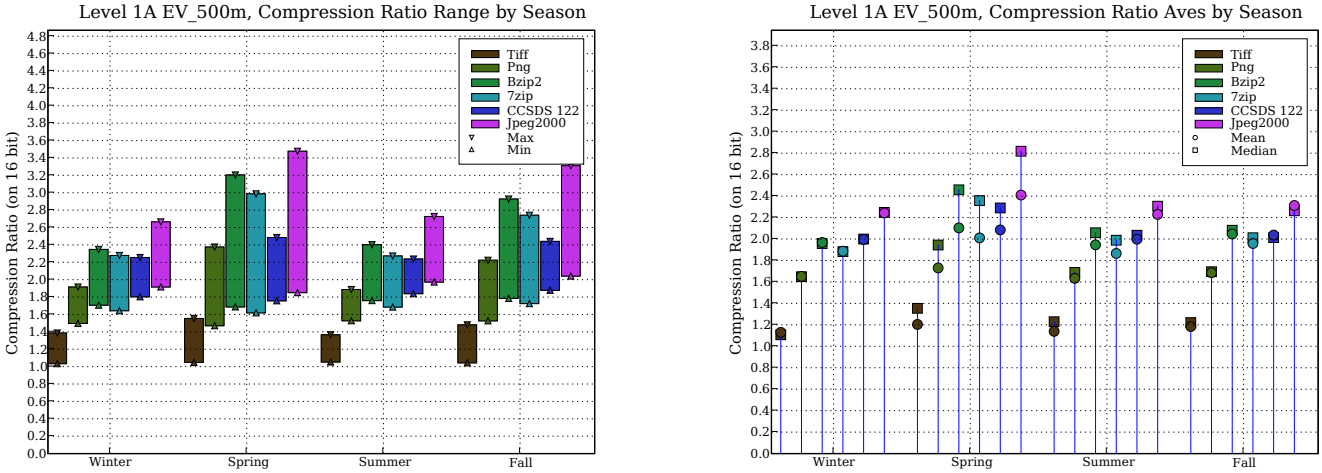

Figure 8. Graphs showing ranges and averages of compression ratios of 500m resolution data for each season (corrected for hemisphere). The results for the 500m resolution data follow the results for the 250m resolution data. The same subtle indications of variability for Spring and Fall are visible.

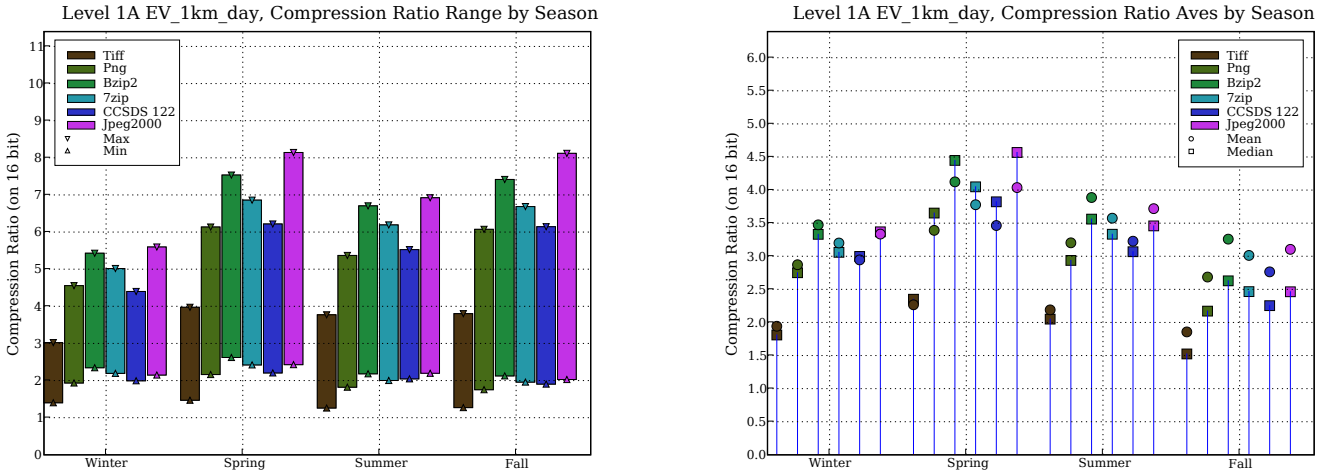

Figure 9. Graphs showing ranges and averages of compression ratios of 1 km resolution reflective data. The ranges show that the Bzip2 algorithm is able to exploit those images which contain little information (due to saturation) in a way that the other algorithms do not. As a result, Bzip2 gives the same or marginally better than Jpeg2000 overall for all seasons.

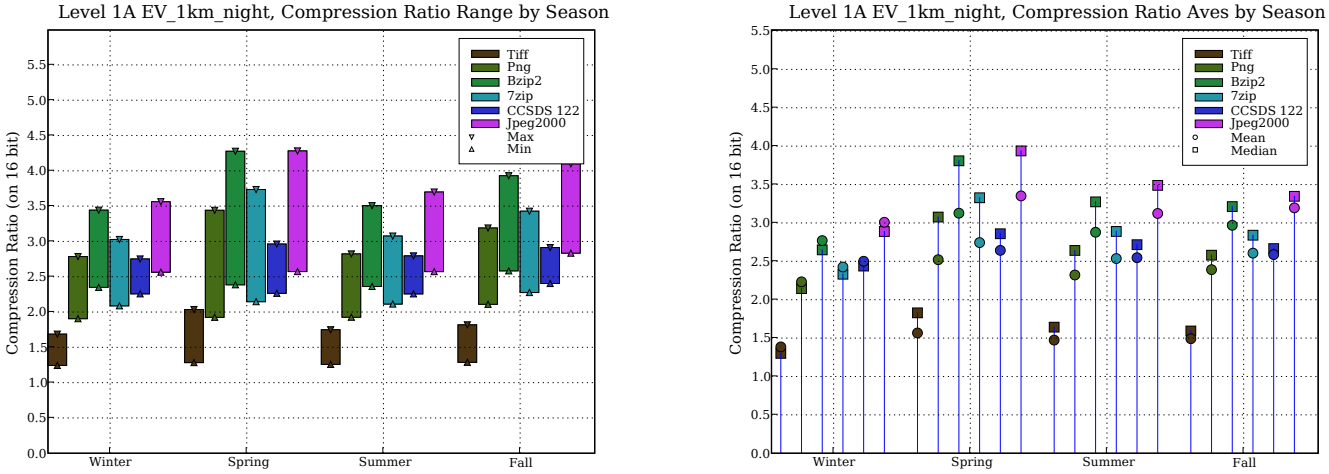

Figure 10. Graphs showing ranges and averages of compression ratios of 1 km resolution em-missive data over the seasons. Jpeg2000 does come out as the overall winner, Bzip2 is quite competitive. The data shows little variation over season.

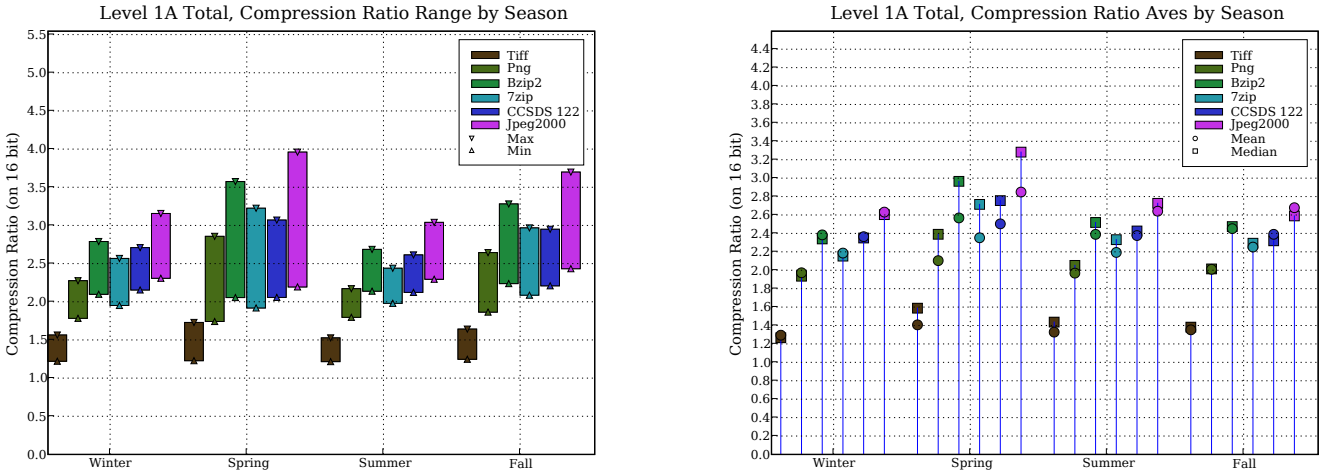

Figure 11. Graphs showing ranges and averages of compression ratios for all the data in a granule considered over the seasons. The variability in the spring data remains clear but all other variation seems to have disappeared. The consistent picture that emerges shows that the performance pattern of Jpeg2000 followed by Bzip2, CCSDS, and 7zip persists.

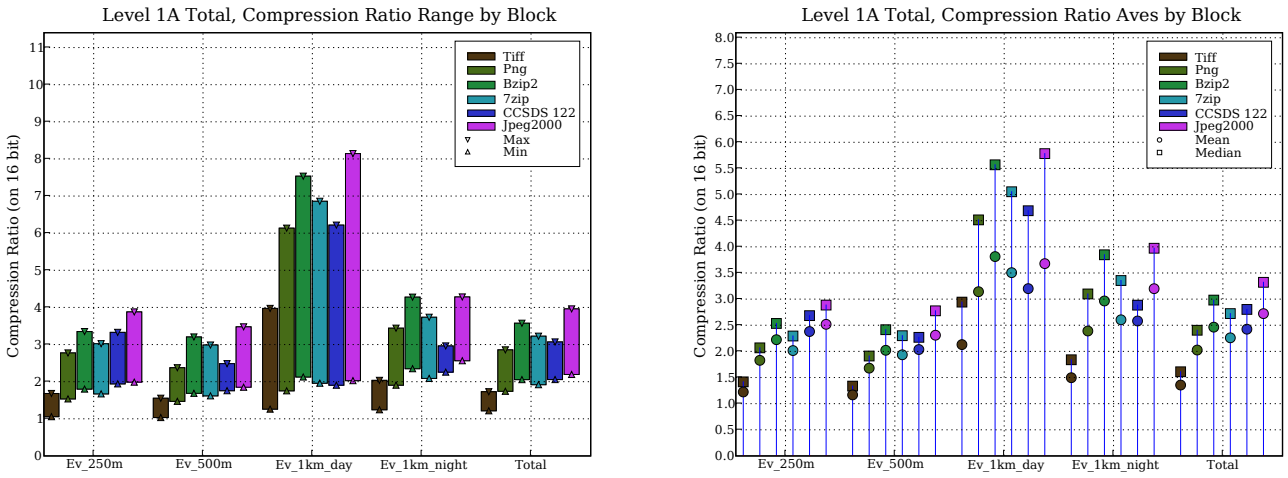

Figure 12. The graphs show the ranges and averages of the compression ratio by data block for all the data in our sample. The 250m an 500m data which make up the bulk of the file size is the most difficult to compress. The 1km reflective data shows the greatest variability, and the 1km night data compresses somewhat better than the other data blocks. Jpeg2000 is the clear winner overall, although Bzip2 does much better than one might expect given that it is not specifically designed for images.

REFERENCES

1. J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding," *IEEE Trans. on Info. Theory* **24**, pp. 531–536, September 1978.
2. T. A. Welch, "A technique for high performance data compression," *IEEE Computer* **17**(6), 1984.
3. I. Pavlov, "http://www.7-zip.org."
4. J. Ziv and A. Lempel, "A Universal algorithm for sequential data compression," *IEEE Trans. on Info. Theory* **23**, pp. 337–343, May 1977.
5. "http://www.bzip.org."
6. P.-S. Yeh, P. Armbruster, A. Kiely, B. Masschelein, G. Moury, C. Schaefer, and C. Thiebaut, "The new CCSDS image compression recommendation," *Aerospace Conference, IEEE* **5-12**, pp. 4138–4145, March 2005.
7. M. D. Adams and F. Kossentini, "JasPer: A software-Based JPEG-2000 Codec Implementation," *Proc. of IEEE International Conference on Image Processing* **2**, pp. 53–56, Oct 2000.